Markov decision processes: dynamic programming and applications

ENSTA Course SOD312 & M2 optimization (Paris-Saclay University and Institut Polytechnique de Paris)

2024 Lecture notes

Marianne Akian

INRIA Saclay - Île-de-France and

CMAP École polytechnique CNRS IP Paris,

marianne.akian@inria.fr

`http://www.cmap.polytechnique.fr/~akian/cours-ensta/`

Abstract

The aim of this course is to introduce different stochastic control models and to present dynamic programming as a tool for solving them. Illustrations selected among stock management, portfolio selection, Yield management, transportation or Web PageRank optimisation will be presented.

We shall consider essentially stochastic dynamical systems with discrete time and finite state space, or finite Markov chains. This framework already contains the essential difficulties (for instance for long term problems), and allows one to give at the same time an insight of algorithms, mathematical techniques and qualitative properties. We may however consider some examples with infinite state space or continuous time.

We shall present the different types of stochastic control problems: complete and incomplete observation problems, criteria with finite horizon, discounted infinite horizon, stopping time, ergodic criteria, risk-sensitive criteria, constrainted problems, armed bandit problems. For some of these criteria, we shall state the corresponding dynamic programming equations, study their qualitative properties, and the algorithms for solving them (value iterations, policy iterations, linear programming), and deduce in some cases the structure of optimal strategies.

Key Words: Markov Decision processes, Stochastic control, Ergodic control, Risk-sensitive control, Dynamic programming, Max-plus algebra, Value iteration, Policy iteration.

Contents

Motivations and Introduction

A *Markov decision problem*, or a *deterministic or stochastic control problem* consists in the maximization or minimization of some functional involving a (possibly random) dynamical system and constructed dynamically. This means that we consider an optimization problem in which the variables are:

- A dynamical system $(X_t)_{t\geq 0}$ over a *state space* $\mathcal{E}$;
- A control process $(U_t)_{t\geq 0}$ taking its values in a *control or action space* $\mathcal{C}$, on which the states depend;
- both may depend on a random process $(W_t)_{t\geq 0}$.

In particular the simplest stochastic control problem satisfies

$$X_{k+1} = f_k(X_k, U_k, W_k), \quad k \in \mathbb{N}, \tag{0.1}$$

where $(W_k)_{k\geq 0}$ is a sequence of independent random variables.

The criteria J to be optimized has a dynamical structure, which "separate" past and future of the state, and thus allows to apply *Dynamic Programming method* (introduced by (Bellman, 53)). For instance, it may be additive:

$$J(X;U) := \sum_{k=0}^{T} g_k(X_k, U_k) \quad \text{(for a discrete time problem)}.$$

Applications

Several real life problems can be modelized as Markov decision processes (MDP) or Stochastic Control Problems. Here are some examples:

- Airline Revenue management;
- Portfolio selection;
- Dam management;
- Stock management;
- Transportation or Web PageRank optimisation;
- Divorce of birds;

Aim of the course

- Modelize real life problems.
- Apply dynamic programming approach.
- Solve dynamic programming equations:
  - with analytical tools (convexity,monotonicity,...)
  - with numerical algorithms (value and policy iterations, linear programming)

One shall consider essentially deterministic or stochastic dynamical systems with discrete time and finite state space, and go from simple to more sophisticated models/problems:

- from deterministic to stochastic problems;
- from uncontrolled to controlled problems;
- from complete to incomplete observation problems;
- additive criteria with finite horizon, discounted infinite horizon, stopping time, ergodic criteria, risk-sensitive criteria;
- from unconstrained to constrainted problems.

References

References [2, 3, 5, 7] contain material similar to the contains of this course. Background material can be found in [4, 1, 6]. Additional references will be given during the lectures.

[1] Robert B. Ash. *Basic probability theory.* John Wiley & Sons, Inc., New York-London-Sydney, 1970.

[2] D. P. Bertsekas. *Dynamic programming.* Prentice Hall Inc., Englewood Cliffs, NJ, 1987.

[3] E.B. Dynkin and A.A. Yushkevich. *Controlled Markov processes.* Springer-Verlag, New York, 1979.

[4] William Feller. *An introduction to probability theory and its applications. Vol. I.* Third edition. John Wiley & Sons, Inc., New York-London-Sydney, 1968.

[5] M. L. Puterman. *Markov decision processes: discrete stochastic dynamic programming.* John Wiley & Sons Inc., New York, 1994.

[6] R. T. Rockafellar. *Convex analysis.* Princeton University Press Princeton, N.J., 1970.

[7] P. Whittle. *Optimization over time. Vol. I and II.* John Wiley & Sons Ltd., Chichester, 1982, 1983.

Chapter 1

Dynamic programming principle for deterministic optimal control

1.1 Dynamical systems (some recalls)

Let us recall what is a general dynamical system.

Definition 1.1. A (deterministic) dynamical system consists in a function (or sequence) from a set $\mathcal{T}$ of *times* to a set $\mathcal{E}$ of *states*, denoted $(X_t)_{t\in\mathcal{T}}$, such that, for each time $t \in \mathcal{T}$, the state X_t of the system at time t is a (deterministic) function f_t of the history of the states until time t, that is $(X_\tau)_{\tau<t}$.

There, the set of times $\mathcal{T}$ may be:

- $\mathbb{Z}$ or $\mathbb{N}$ and $t = n$: we speak about *discrete time* dynamical system.
- $\mathbb{R}$ or $\mathbb{R}_+$: we speak about *continuous time* dynamical system.

The state space $\mathcal{E}$ may be:

- a finite set: *finite state space*.
- a finite or countable set: *discrete state space*.
- $\mathbb{R}^n$: (finite dimensional) *continuous state space* /system.
- a space of functions : *infinite dimensional continuous state space* /system.

The system may satisfy:

- $X_n = f_n(X_{n-1})$, $n \geq 0$.
- $\dot{X} = g_t(X)$, $t \geq 0$, with g_t Lipschitz continuous: mechanical or physical systems.
- $\frac{\partial X}{\partial t} = -\Delta X$. Heat equation.
- the discretization of a ODE or PDE.

In particular,

Definition 1.2. A (deterministic) dynamical system with discrete time and state space $\mathcal{E}$ is a sequence $(X_k)_{k\in\mathbb{Z}}$ or $(X_k)_{k\in\mathbb{N}}$ with values in the set $\mathcal{E}$, such that, for each $k \in \mathbb{N}$, the state X_k at *time (or stage)* k is a (deterministic) function f_k of the states at previous times, that is of $X_{k-1}, X_{k-2}, \ldots$.

The sequence $(X_k)_{k\geq 0}$ is called the *trajectory* of the system starting from X_0.

Examples of discrete time dynamical systems are as follows:

1. $X_{n+1} = f_n(X_n)$, $n \geq 0$, where $f_n : \mathcal{E} \to \mathcal{E}$.
2. $X_{n+1} = f_n(X_n, X_{n-1})$, $n \geq 1$, with $f_n : \mathcal{E} \times \mathcal{E} \to \mathcal{E}$.
3. $X_{n+1} = f_n(X_n, X_{n-1}, \ldots, X_0)$, $n \geq 0$, with $f_n : \mathcal{E}^{n+1} \to \mathcal{E}$.
4. $X_{n+1} = f_n(X_{n-\tau(n)})$, $n \geq 0$, where $\tau : \mathbb{N} \to \mathbb{N}$ is a variable delay.

Fact 1.3. Any discrete time dynamical system can be reduced to a system of type 1.

Proof. If the time set is $\mathbb{Z}$, then a dynamical system is such that $X_{n+1} = f_n(X_n, X_{n-1}, \ldots)$, with $f_n : \mathcal{E}^{\mathbb{N}} \to \mathcal{E}$, for all $n \in \mathbb{Z}$.

Consider the new state $X'_k = (X_k, X_{k-1}, \ldots)$ belonging to the larger state space $\mathcal{E}' = \mathcal{E}^{\mathbb{N}}$, then X'_k has the dynamics $X'_{k+1} = f'_k(X'_k)$ with

$$f'_k((x_0, x_1, \ldots)) = (f_k(x_0, x_1, \ldots), x_0, x_1, \ldots) \ .$$

If the time set is $\mathbb{N}$, a dynamical system is such that $X_{n+1} = f_n(X_n, X_{n-1}, \ldots, X_0)$, with $f_n : \mathcal{E}^{n+1} \to \mathcal{E}$, for all $n \geq 0$.

We can reduce the new state space to $\mathcal{E}' = \cup_{k\geq 0}\mathcal{E}^{k+1}$, which is countable if $\mathcal{E}$ is a finite set. Indeed, the new state $X'_k = (X_k, X_{k-1}, \ldots, X_0)$ belongs to $\mathcal{E}'$ and has the dynamics $X'_k = f'_k(X'_{k-1})$, where f'_k is only defined on $\mathcal{E}^k$ by

$$f'_k((x_0, \ldots, x_{k-1})) = (f_k(x_0, \ldots, x_{k-1}), x_0, x_1, \ldots, x_{k-1}) \in \mathcal{E}^{k+1} \ .$$

□

A drawback of the above construction is that even if $\mathcal{E}$ were a finite set, the new state space $\mathcal{E}'$ may be *infinite noncountable*, in particular, when the time set is $\mathbb{Z}$. Also to initialize the sequence, one need an initial state of the form $X'_0 = (X_0, X_{-1}, \ldots)$ to be given. However, when the time set is $\mathbb{N}$, and $\mathcal{E}$ is a finite set, then the state space $\mathcal{E}'$ is countable.

The above construction is similar to the one used to transform a second order differential equation to a first order one for instance.

1.2 Deterministic optimal control problems with additive payoff and finite horizon

The simplest deterministic control problem is the following. Consider a discrete time dynamical system $(X_n)_{n\geq 0}$ with finite (or discrete) state space $\mathcal{E}$ and a dynamics of type 1:

$$X_{n+1} = f_n(X_n), \quad n \geq 1 \ .$$

Assume now that we (or somebody) are able to change the behavior of this system, that is its dynamics f_n.

That is, the dynamics is supposed to depend not only on the state but also on a parameter, called the *action* or the *control*:

$$X_{n+1} = f_n(X_n, U_n), \quad n \geq 1 \ .$$

An *optimal control problem* is the problem of choosing the actions $U_0, \ldots, U_k, \ldots$ in such a way that they minimize (resp. maximize) a certain functional, called the total cost (resp. the total payoff) of the sequences $X = (X_k)_{k\geq 0}$ and $U = (U_k)_{k\geq 0}$.

We assume in this part that all the states are observable, which means in particular that the initial state X_0 is known. We speak about *complete observation*.

Let us consider or denote:

- a finite or discrete *state space* $\mathcal{E}$;
- an *action space* $\mathcal{C}$
- for all $k \in \mathbb{N}$ and $x \in \mathcal{E}$, the subset $\mathcal{C}_k(x) \subset \mathcal{C}$ of all possible actions at time k, when the state is equal to x;
- for all $k \in \mathbb{N}$, the set $\mathcal{A}_k := \{(x, u) \mid x \in \mathcal{E},\ u \in \mathcal{C}_k(x)\}$ of all possibles couples (state, action) at time k;
- for all $k \geq 0$, the *dynamics* at time k, which is a map $f_k : \mathcal{A}_k \to \mathcal{E}$;
- for all $k \in \mathbb{N}$, the *instantaneous/running reward/payoff* at time k, which is a map $r_k : \mathcal{A}_k \to \mathbb{R}$;
- a *final reward*, which is a map $\varphi : \mathcal{E} \to \mathbb{R}$;
- an *initial state* $x_0 \in \mathcal{E}$.
- for all sequences $X = (X_k)_{k\geq 0}$ and $U = (U_k)_{k\geq 0}$ in $\mathcal{E}$ and $\mathcal{C}$ respectively, the *total additive payoff* with finite horizon $T \geq 1$:

$$J(X; U) := \left(\sum_{k=0}^{T-1} r_k(X_k, U_k) \right) + \varphi(X_T) \ . \tag{1.1}$$

Moreover, we shall replace the words *reward* or *payoff* by *cost*, and the notation r_k by c_k, when the criterion J is to be *minimized*, instead of maximized.

Definition 1.4. A deterministic control problem with discrete time, complete observation, and the above data and additive criteria consists in the following optimization problem:

$$\max_{X,U} J(X; U)$$

where the optimization holds over all sequences $X = (X_k)_{k\geq 0}$ and $U = (U_k)_{k\geq 0}$ in $\mathcal{E}$ and $\mathcal{C}$ respectively, such that

$$X_{k+1} = f_k(X_k, U_k),\ X_0 = x_0,\ U_k \in \mathcal{C}_k(X_k),\ k \in \mathbb{N} \ .$$

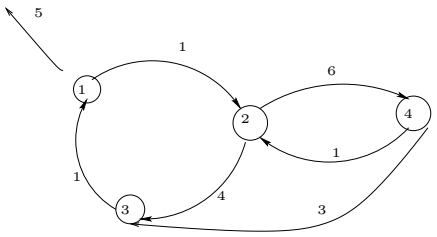
The optimum of above criteria is called the *value* of the problem. A sequence $U = (U_k)_{k\geq 0}$ of an optimal solution (X, U) is called an *optimal control (process)*. Moreover, maximization can be replaced by minimization.

Definition 1.5. For all $x_0 \in \mathcal{E}$, let $v(x_0)$ be the value of the problem of Definition 1.4 when the initial state is x_0. The map $v : \mathcal{E} \to \mathbb{R}, x_0 \mapsto v(x_0)$ is called the *value function*.

Some small generalizations of the above problem can be done.

- One can also consider functions r_k and φ taking their values in $\mathbb{R} \cup \{-\infty\}$ (for a maximization problem).
- This allows one to restrict the state space at each time. Indeed, for all $k \geq 0$, let $\mathcal{E}_k \subset \mathcal{E}$, and take $r_k(x, u) = -\infty$ when $x \notin S_k$ and $\varphi(x) = -\infty$ when $x \notin \mathcal{E}_T$, then the maximum of J is attained only for a sequence X such that $X_k \in \mathcal{E}_k$ for all $k \geq 0$.
- In particular, if $\varphi(x) = 0$ when $x = x_T$ and $\varphi(x) = -\infty$ otherwise, then the final state is necessarily equal to x_T.
- Conversely, one can replace the constraint $X_0 = x_0$ by the addition to J of an initial reward $\psi : \mathcal{E} \to \mathbb{R} \cup \{-\infty\}$, as in
$$\max_{X,U} \psi(X_0) + J(X; U)$$
where the optimization holds over all sequences $X = (X_k)_{k\geq 0}$ and $U = (U_k)_{k\geq 0}$ in $\mathcal{E}$ and $\mathcal{C}$ respectively, such that
$$X_{k+1} = f_k(X_k, U_k),\ U_k \in \mathcal{C}_k(X_k),\ k \in \mathbb{N}\ .$$
- If $\psi(x) = 0$ when $x = x_0$ and $\varphi(x) = -\infty$ otherwise, we recover the previous problem.
- If v_ψ denotes the value of the problem, then
$$v_\psi = \max_{x_0 \in \mathcal{E}} (\psi(x_0) + v(x_0))\ .$$

Example 1.6 (**Shortest path problem).** Consider a directed graph $\mathcal{G} = (\mathcal{N}, \mathcal{A})$, where $\mathcal{N}$ is the set of nodes, and $\mathcal{A} \subset \mathcal{N}^2$ is the set of arcs. Let ℓ be a weight function representing the "lengths" of arcs: $\ell : \mathcal{A} \to \mathbb{R}_+$. One can think for instance to a network of towns: the nodes are the towns, the arcs the direct roads between some of them (that is the ones containing no other town), and ℓ can be either the true length (in km) of the road or the gas consumption necessary to travel on it. The shortest path problem starting from node x_0 and ending in x_f consists in solving:
$$\min \left\{ \sum_{k=0}^{N-1} \ell(x_k, x_{k+1}) \mid N \in \mathbb{N},\ (x_0, \ldots, x_N) \text{ is a path of } \mathcal{G},\ x_N = x_f \right\}\ .$$



When we restrict the minimization to paths with length (that is number of arcs) less or equal to N, we get:

$$\min \left\{ \sum_{k=0}^{n-1} \ell(x_k, x_{k+1}) \mid n \leq N,\ (x_0, \ldots, x_n) \text{ is a path of } \mathcal{G},\ x_n = x_f \right\} .$$

This problem is an optimal control problem with finite horizon N with

- $\mathcal{E} = \mathcal{C} = \mathcal{N}$.
- $\mathcal{C}_k(x)$ is the set of nodes y such that $(x, y) \in \mathcal{A}$, when $x \neq x_f$ and $\mathcal{C}_k(x_f) = \{x_f\}$.
- The dynamics is $f_k(x, u) = u$.
- The instantaneous cost is : $r_k(x, u) = \ell(x, u)$ if $x \neq x_f$ and $r_k(x_f, x_f) = 0$.
- The final cost is $\varphi(x) = 0$ if $x = x_f$ and $\varphi(x) = +\infty$ otherwise.
- The initial state is x_0.

Example 1.7 (**Ressource allocation problem).** An investor can invest $M \in \mathbb{N}$ units (of money, capacity, requets,...) in $N \in \mathbb{N}$ ressources (stocks, flats, planes, parallel processors,...).

- We assume that the reward obtained when he invests x units in the ith ressource is equal to $R_i(x)$.
- We also assume that the units that are not invested yield a zero reward.
- So the investor wants to maximize his total reward, that is equivalent to find

$$\max \left\{ \sum_{i=1}^{N} R_i(u_i) \mid u_i \in \mathbb{N},\ i = 1, \ldots, N,\ \sum_{i=1}^{N} u_i \leq M \right\} .$$

Consider the deterministic optimal control problem in which:

- The ressource number is considered as a stage/time;
- The state at each stage is equal to the number of units remaining to be invested.
- The state space is $\mathcal{E} = \{0, \ldots, M\}$;
- The action spaces are $\mathcal{C} = \mathcal{E}$ and $\mathcal{C}_k(x) = \{0, \ldots, x\}$.
- The dynamics at each time k is : $f_k(x, u) = x - u$;
- for all $k \in \mathbb{N}$, the instantaneous reward is : $r_k(x, u) = R_{k+1}(u)$.
- The final reward is $\varphi(x) = 0$.
- The initial state is $x_0 = M$.

- The total additive payoff with finite horizon $T = N$ is then:

$$J(X;U) = \sum_{k=0}^{T-1} r_k(X_k, U_k) = \sum_{k=0}^{N-1} R_{k+1}(U_k) \ .$$

Then the value of this problem coincides with the one of the ressource allocation problem : take $u_k = U_{k-1}$ and X_k equal to the number of units remaining to be invested, when the number of units $u_1, \dots, u_k$, invested in ressources 1 to k have already been chosen ($X_k = X_0 - u_1 - \cdots - u_k$).

Example 1.8 (Knapsack problem). Given N items, each with a weight w_i and a value m_i, $i = 1, \dots, N$, one need to determine the number u_i of each item i to include in a knapsack so that the total weight is less than or equal to W and the total value is as large as possible. This consists in the optimization problem:

$$\max \left\{ \sum_{i=1}^{N} m_i u_i \mid u_i \in \mathbb{N},\ i = 1, \dots, N,\ \sum_{i=1}^{N} w_i u_i \leq W \right\} \ .$$

Additionnaly, the numbers u_i can be restricted to be in $\{0, 1\}$ or to be in $\{0, \dots, c\}$.

If $w_i \in \mathbb{N}^*$, for all $i = 1, \dots, N$, considering the ressource allocation problem in which the rewards are linear with $R_i(x) = \frac{m_i}{w_i} x$, and adding constraints induced by non-divisibility of ressources, or restricting the sets $\mathcal{C}_i(x)$ of the optimal control problem, we recover all types of knapsack problems. For instance, for 0-1 knapsack problem, one can consider $\mathcal{C}_{i-1}(x) = \{0, \dots, x\} \cap \{0, w_i\}$. For the unbounded knapsack problem, one can consider $\mathcal{C}_{i-1}(x) = \{0, \dots, x\} \cap w_i \mathbb{N}$.

1.3 Dynamic programming equation

In the optimal control problem of Definition 1.4:

$$\max_{X,U} J(X;U) \tag{1.2a}$$

$$X_{k+1} = f_k(X_k, U_k),\ X_0 = x_0,\ U_k \in \mathcal{C}_k(X_k),\ k \in \mathbb{N} \tag{1.2b}$$

the optimization is done over all sequences $U = (U_k)_{k \geq 0}$ satisfying the above constraints.

One may wish to obtain an optimal control which yields a dynamical system, that is a system which is causal in the sense that the state X_{k+1} at time $k+1$ only depends on the past states $X_0, \dots, X_k$. Moreover, one may wish to take a decision at time k using only the informations we have in hand, that is the history of the trajectories of X and U before the decision. A *strategy* is precisely a rule which tells how to take this decision.

Definition 1.9. The set $\mathcal{H}_k = \mathcal{A}_0 \times \cdots \times \mathcal{A}_{k-1} \times \mathcal{E}$ is called the set of *histories* at time k.

A *strategy* for the (perfect information) optimal control problem of Definition 1.4 is a sequence $\sigma = (\sigma_0, \dots, \sigma_{T-1})$ such that, for all $k = 0, \dots, T-1$, σ_k is a map from $\mathcal{H}_k$ to $\mathcal{C}$ satisfying

$$\sigma_k(x_0, u_0, \dots, x_{k-1}, u_{k-1}, x_k) \in \mathcal{C}_k(x_k), \text{ for all } (x_0, u_0, \dots, x_{k-1}, u_{k-1}, x_k) \in \mathcal{H}_k \ .$$

We denote by $\Sigma^{(T)}$ the set of all strategies. A strategy gives rise to a dynamical system $(X_k, U_k)_{k \geq 0}$ satisfying the dynamics: $X_{k+1} = f_k(X_k, U_k)$ and $U_k = \sigma_k(X_0, U_0, \dots, X_{k-1}, U_{k-1}, X_k)$. Such a sequence $(X_k, U_k)_{k \geq 0}$ is also called an *admissible sequence* of states and controls.

A strategy is *optimal* if the sequence (X, U) is optimal for the optimization problem (1.2) restricted to admissible sequences of states and controls (that is to strategies).

Definition 1.10. A strategy σ is a *feedback policy* if each map σ_k depends only on the information on the state at the current time, that is

$$\sigma_k(x_0, u_0, \dots, x_{k-1}, u_{k-1}, x_k) = \pi_k(x_k) \ ,$$

where $\pi_k : \mathcal{E} \to \mathcal{C}$ is such that $\pi_k(x_k) \in \mathcal{C}_k(x_k)$. We then denote by $\pi = (\pi_0, \dots, \pi_{T-1})$ such a policy, and by $\Pi^{(T)}$ the set of all feedback policies.

Definition 1.11. An *open-loop control* is a strategy such that each map σ_k depends only on the state x_0 at the initial time, that is

$$\sigma_k(x_0, u_0, \dots, x_{k-1}, u_{k-1}, x_k) = \omega_k(x_0) \ .$$

We denote by $O^{(T)}$ the set of all open-loop controls.

The following result shows that the optimization of the total reward J over each of the above types of strategies gives the same value. However, one can show that feedback policies are more robust with respect to disturbances on the model, that is disturbances on the maps f_k and r_k.

Theorem 1.12 ((Bellman) Dynamic programming method for deterministic optimal control problems)**.** *Assume that the maps φ, r_k, $k \geq 0$ are bounded from above. Define the functions $v_t : \mathcal{E} \to \mathbb{R}$, $t = 0, \dots, T$, by the* backward recursion*:*

$$v_T(x) = \varphi(x) \quad \forall x \in \mathcal{E} \ , \tag{1.3a}$$

$$v_k(x) = \sup\{r_k(x,u) + v_{k+1}(f_k(x,u)) \mid u \in \mathcal{C}_k(x)\} \quad \forall x \in \mathcal{E},\ k \leq T-1. \tag{1.3b}$$

Then the value function v of the optimal control problem of Definition 1.4 coincides with v_0.

Moreover, the value v coincides with the optimum of (1.2) *(that is of J) restricted to admissible controls (or strategies), or to feedback policies, or to open-loop controls.*

Assume in addition that the maximum of (1.3b)*, is attained for an action $u \in \mathcal{C}_k(x)$ and* let us denote by $\pi_k(x)$ this action*, then the feedback policy $\pi = (\pi_k)_{0 \leq k \leq T-1}$ is an optimal strategy of the problem, and the dynamics $X_{k+1} = f_k(X_k, \pi_k(X_k))$ with $U_k = \pi_k(X_k)$ furnishes an optimal solution (X, U) of the optimal control problem.*

Proof. The dynamic programming equation by moving suprema. The value $v(x_0)$ of the problem of Definition 1.4 is by definition the optimum of

$$J(X;U) := \left(\sum_{k=0}^{T-1} r_k(X_k, U_k)\right) + \varphi(X_T)$$

over all sequences $X = (X_k)_{k \geq 0}$ and $U = (U_k)_{k \geq 0}$ of $\mathcal{E}$ and $\mathcal{C}$ satisfying the constraints (1.2b):

$$v(x_0) = \sup\left\{ \left(\sum_{k=0}^{T-1} r_k(X_k, U_k)\right) + \varphi(X_T) \mid (X,U) \text{ satisfies (1.2b)} \right\} \ .$$

Note that these sequences are not necessarily admissible.

This can be rewritten as:

$$v(x_0) = \sup_{U_0 \in \mathcal{C}_0(x_0),\, X_1 = f_0(x_0,U_0)} \left(\cdots \sup_{U_{T-1} \in \mathcal{C}_{T-1}(X_{T-1}),\, X_T = f_{T-1}(X_{T-1},U_{T-1})} \left(\left(\sum_{k=0}^{T-2} r_k(X_k,U_k) \right) + r_{T-1}(X_{T-1},U_{T-1}) + \varphi(X_T) \right) \cdots \right) \ ,$$

Consider the maps v_k, $k \geq 0$, as in (1.3).

Since $r_0(X_0,U_0), \ldots, r_{T-2}(X_{T-2},U_{T-2})$ do not depend on U_{T-1} nor X_T, but only on the first states and actions $X_0, \ldots, X_{T-2}$ and $U_0, \ldots, U_{T-2}$, we deduce that for $X_0 = x_0$, we have

$$\begin{aligned} v(x_0) &= \sup_{U_0 \in \mathcal{C}_0(X_0),\, X_1 = f_0(X_0,U_0)} \left(\cdots \sup_{U_{T-2} \in \mathcal{C}_{T-2}(X_{T-2}),\, X_{T-1} = f_{T-2}(X_{T-2},U_{T-2})} \right. \\ &\quad \left(\left(\sum_{k=0}^{T-2} r_k(X_k,U_k) \right) + \right. \\ &\quad \left. \left. \sup_{U_{T-1} \in \mathcal{C}_{T-1}(X_{T-1})} \left(r_{T-1}(X_{T-1},U_{T-1}) + v_T(f_{T-1}(X_{T-1},U_{T-1})) \right) \right) \cdots \right) \ , \\ &= \sup_{U_0 \in \mathcal{C}_0(X_0),\, X_1 = f_0(X_0,U_0)} \left(\cdots \sup_{U_{T-2} \in \mathcal{C}_{T-2}(X_{T-2}),\, X_{T-1} = f_{T-2}(X_{T-2},U_{T-2})} \right. \\ &\quad \left. \left(\left(\sum_{k=0}^{T-2} r_k(X_k,U_k) \right) + v_{T-1}(X_{T-1}) \right) \cdots \right) \ , \\ &= \cdots = v_0(X_0) \ . \end{aligned}$$

which shows that the value function v of the problem of Definition 1.4 coincides with the function v_0 defined recursively by (1.3).

Optimality and an alternative proof of dynamic programming equation Applying (1.3b) recursively, it is easy to show that, for all sequences (X,U) of states and actions satisfying the dynamics $X_{k+1} = f_k(X_k,U_k)$, we have

$$v_0(x_0) \geq r_0(x_0,U_0) + v_1(X_1) \geq \cdots \geq J(X;U) \ .$$

Taking the supremum over all sequences, we deduce that $v_0(x_0) \geq v(x_0)$.

Now, if $\pi_k(x)$ is optimal in the criteria (1.3b), then taking the sequence (X,U) such that $U_k = \pi_k(X_k)$ and $X_{k+1} = f_k(X_k,U_k)$, we get $v_0(x_0) = r_0(X_0,U_0) + v_1(X_1) = \cdots = J(X;U) \leq v(x_0)$. Hence, since $v_0(x_0) \geq v(x_0)$, we deduce the equality and that (X,U) is optimal.

Moreover, this action is a function of k and X_k, hence it comes from a (feedback) strategy, which is in particular a strategy. This shows that the maximum $v(x_0)$ over all sequences is equal to the maximum over all feedback strategies, or over all strategies.

Since the dynamics is deterministic, the constraints $U_k = \pi_k(X_k)$ together with (1.2b) allow to write U_k as a function of X_0 only, that is as an open-loop control, hence we also get that the maximum $v(x_0)$ coincides with the optimum of J restricted to all open-loop controls.

When the action sets are infinite, but the suprema are finite, which is the case when the maps φ, r_k, $k \geq 0$, are bounded from above, one can prove the same result by considering actions $\pi_k(x)$ that are ϵ-optimal for (1.3b), that is satisfying for all $x \in \mathcal{E}$, and $k \leq T-1$,

$$v_k(x) \leq \epsilon + r_k(x, \pi_k(x)) + v_{k+1}(f_k(x, \pi_k(x))) \ .$$

Indeed, this gives a control sequence $U = (\pi_k(X_k))_{k \geq 0}$ which is $(T\epsilon)$-optimal for the criteria J :

$$v_0(x_0) \leq \epsilon + r_0(x_0, U_0) + v_1(X_1) \leq \cdots \leq T\epsilon + J(X;U) \leq T\epsilon + v(x_0) \ .$$

Since this holds for all $\epsilon > 0$, we obtain that $v_0(x_0) \leq v(x_0)$ and so the equality as in the above case of finite action sets. Moreover, we obtain that $v(x) \leq v_0(x_0) \leq T\epsilon + J(X;U)$ and since the sequence (X, U) comes from a feedback strategy, this shows that the supremum of $J(X;U)$ over all feedback strategies is equal to $v(x_0)$. The other assertions are shown as for the case of finite action sets. □

Remark 1.13. Let us consider the partial criteria:

$$J_n(X;U) := \left(\sum_{k=n}^{T-1} r_k(X_k, U_k) \right) + \varphi(X_T)$$

Then, the iterations $(v_n)_{0 \leq n \leq T}$ defined in the dynamic programming equation have the following interpretation:

$$v_n(x_n) = \max_{X,U} J_n(X;U) \tag{1.4a}$$

where the optimization is done over all sequences $X = (X_k)_{k \geq n}$ and $U = (U_k)_{k \geq n}$ satisfying the following constraints

$$X_{k+1} = f_k(X_k, U_k),\ X_n = x_n,\ U_k \in \mathcal{C}_k(X_k),\ n \leq k \leq T-1 \tag{1.4b}$$

Example 1.14 (**Shortest path problem (continued)).** Let us consider the shortest path problem described in Example 1.6. Using the optimal control interpretation of the value of the shortest path from x_0 to x_f with paths with length (that is number of arcs) less or equal to N:

$$v^{(N)}(x_0) := \min \left\{ \sum_{k=0}^{n-1} \ell(x_k, x_{k+1}) \mid n \leq N,\ (x_0, \ldots, x_n) \text{ is a path of } \mathcal{G},\ x_n = x_f \right\} \ ,$$

we obtain the dynamic programming equation:

$$\begin{aligned}
&v_N^{(N)}(x) = +\infty \quad \forall x \in \mathcal{N} \setminus \{x_f\} \ , \\
&v_N^{(N)}(x_f) = 0 \ , \\
&v_k^{(N)}(x) = \min\{\ell(x,y) + v_{k+1}^{(N)}(y) \mid y,\ (x,y) \in \mathcal{A}\} \quad \forall x \in \mathcal{N} \setminus \{x_f\},\ k \leq N-1 \ , \\
&v_k^{(N)}(x_f) = v_{k+1}^{(N)}(x_f) \ .
\end{aligned}$$

Morover, since ℓ does not depend on time k, we can rewrite theses equation as a *forward* recurrence for $w_k = v_0^{(k)}$, such that $v_k^{(N)} = w_{N-k}$:

$$\begin{aligned}
&w_0(x) = +\infty \quad \forall x \in \mathcal{N} \setminus \{x_f\} \ , \\
&w_0(x_f) = 0 \ , \\
&w_k(x) = \min\{\ell(x,y) + w_{k-1}(y) \mid y,\ (x,y) \in \mathcal{A}\} \quad \forall x \in \mathcal{N} \setminus \{x_f\},\ k \leq N-1 \ , \\
&w_k(x_f) = w_{k-1}(x_f) \ .
\end{aligned}$$

Example 1.15 (**Ressource allocation problem (continued)).** Let us consider the ressource allocation problem described in Example 1.7. Using the deterministic optimal control problem interpretation, the value of the ressource allocation problem coincides with $v_0(M)$, where v_t, $t = 0, \ldots, N$ satisfy the dynamic programming equation:

$$v_N(x) = 0 \quad \forall x \in \{0, \ldots, M\} \ ,$$
$$v_k(x) = \sup\{R_{k+1}(u) + v_{k+1}(x-u)) \mid 0 \leq u \leq x\} \quad \forall 0 \leq x \leq M,\ k \leq N-1.$$

Example 1.16 (**Knapsack problem (continued)).** Let us consider Knapsack problem described in Example 1.8:

$$\max \left\{ \sum_{i=1}^{N} m_i u_i \mid u_i \in \{0,1\},\ i = 1, \ldots, N,\ \sum_{i=1}^{N} w_i u_i \leq W \right\} \ .$$

As for the ressource allocation problem, the value of the problem is equal to $v_0(W)$ where v_t, $t = 0, \ldots, N$ satisfy the dynamic programming equation:

$$v_N(x) = 0 \quad \forall x \in \{0, \ldots, W\} \ ,$$
$$v_k(x) = \sup\{m_{k+1}u + v_{k+1}(x - w_{k+1}u)) \mid u \in \{0,1\},\ w_{k+1}u \leq x\} \quad \forall 0 \leq x \leq W,\ k \leq N-1.$$

So the recurrence equation reduces to:

$$v_k(x) = \max(v_{k+1}(x), m_{k+1} + v_{k+1}(x - w_{k+1})) \tag{1.5}$$

if $x \geq w_{k+1}$, and $v_k(x) = v_{k+1}(x)$ otherwise. Moreover, one can use (1.5) for all $x \geq 0$, by extending the functions v_k by $v_k(x) = -\infty$ for all negative integers x.

Exercise 1.3.1. Solve the Knapsack problem

$$\max \{4u_1 + 3u_2 + 2u_3 \mid u_i \in \{0,1\},\ i = 1, \ldots, 3,\ 5u_1 + 4u_2 + 3u_3 \leq 10\} \ ,$$

using dynamic programming equation.

1.4 Properties of Dynamic programming

1.4.1 Complexity

Corollary 1.17 (of Dynamic programming)**.** *Under the assumptions of Theorem 1.12, denote, for $x, y \in \mathcal{E}$, and $k \geq 0$:*

$$G_k(x,y) = \sup\{r_k(x,u) \mid u \in \mathcal{C}_k(x),\ f_k(x,u) = y\} \in \mathbb{R} \cup \{-\infty\} \ .$$

Then, the dynamic programming equation of Theorem 1.12 can be rewritten as

$$v_k(x) = \sup\{G_k(x,y) + v_{k+1}(y) \mid y \in \mathcal{E}\} \quad \forall x \in \mathcal{E},\ k \leq T-1.$$

Note also that an optimal feedback policy $\pi_k(x)$ can be obtained as the composition: $\pi_k(x) = \pi'_k(x, \pi''_k(x))$ where

$$\pi'_k(x,y) \in \mathrm{Argmax}\{r_k(x,u) \mid u \in \mathcal{C}_k(x),\ f_k(x,u) = y\}$$

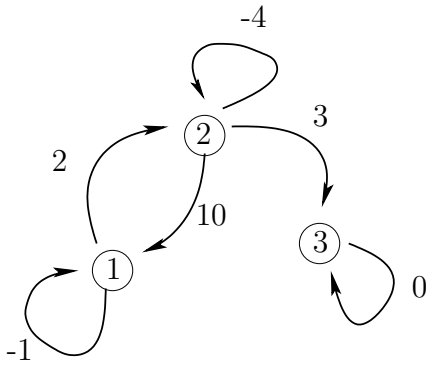
and

$$\pi''_k(x) \in \mathrm{Argmax}\{G_k(x,y) + v_{k+1}(y) \mid y \in \mathcal{E}\}$$

If the maps G_k do not depend on k, one can construct a weighted directed graph with set of nodes $\mathcal{E}$, an arc (x, y) if $G(x, y) \neq -\infty$, with the weight $G(x, y)$. For instance for $\mathcal{E} = \{1, 2, 3\}$ and the following table of values of G

$$G = \begin{bmatrix} -1 & 2 & -\infty \\ 10 & -4 & 3 \\ -\infty & -\infty & 0 \end{bmatrix} ,$$

we obtain the (di)graph:



Then, the weight of a path is the sum of the weights of its arcs, and an optimal sequence of states $(X_k)_{k\geq 0}$ is a path in this graph which has a maximal weight among the paths with same initial and final nodes. When the weights are nonpositive, this is a "shortest path" (for the lengths which are the opposite of G). In general, denote

- $n = \text{card}(\mathcal{E})$, the number of states;
- $m_k = \text{card}(\{(x, y) \in \mathcal{E} \times \mathcal{E} \mid G_k(x, y) \neq -\infty\})$, the number of arcs in the graph.

Fact 1.18. Once an oracle is available to compute G_k, the computational complexity of Dynamic programming equation is $O(\sum_k m_k) = O(Tn^2)$. This has to be compared with $O(n^T)$, if we solve the optimization directly, that is we solve the combinatorial optimization problem consisting in optimizing the criterion over all trajectories $(x_0, \ldots, x_T)$ in $\mathcal{E}^{T+1}$.

The storage complexity is $O(\sum_k m_k)$ if G_k depends on k and is $O(m + Tn)$ otherwise.

1.4.2 Operator properties

Definition 1.19. For $k \leq T - 1$, let $\mathcal{B}_k : \mathbb{R}^{\mathcal{E}} \to \mathbb{R}^{\mathcal{E}}$ be the map such that for all $v \in \mathbb{R}^{\mathcal{E}}$, and $x \in \mathcal{E}$, we have

$$\begin{aligned}[\mathcal{B}_k(v)](x) &= \sup\{r_k(x, u) + v(f_k(x, u)) \mid u \in \mathcal{C}_k(x)\} \\ &= \sup\{G_k(x, y) + v(y) \mid y \in \mathcal{E}\} \ .\end{aligned}$$

The map $\mathcal{B}_k$ is called the *Bellman operator* at time k of the optimal control problem.

The dynamic programming equation can then be rewritten in functional form:

$$v_T = \varphi, \quad v_k = \mathcal{B}_k(v_{k+1}), \quad \text{for } T - 1, \ldots, 0.$$

Definition 1.20. Denotes by $\leq$ the partial order on $\mathbb{R}^{\mathcal{E}}$: $v \leq w$ if $v(x) \leq w(x)$ for all $x \in \mathcal{E}$.

We say that an operator $\mathcal{B} : \mathbb{R}^{\mathcal{E}} \to \mathbb{R}^{\mathcal{E}}$ is *monotone* or *order preserving* if it preserves the partial order of $\mathbb{R}^{\mathcal{E}}$, that is if, for all $v, w \in \mathbb{R}^{\mathcal{E}}$, we have

$$v \leq w \Rightarrow \mathcal{B}(v) \leq \mathcal{B}(w) \ .$$

Definition 1.21. Denote by $\mathbf{1}$ the element of $\mathbb{R}^{\mathcal{E}}$ which is the constant function (or vector) equal to 1: $\mathbf{1}(x) = 1$ for all $x \in \mathcal{E}$.

We say that an operator $\mathcal{B} : \mathbb{R}^{\mathcal{E}} \to \mathbb{R}^{\mathcal{E}}$ is *additively homogeneous* if it commutes with the addition of a constant, that is, for all $v \in \mathbb{R}^{\mathcal{E}}$ and $\lambda \in \mathbb{R}$, we have:

$$\mathcal{B}(v + \lambda\mathbf{1}) = \mathcal{B}(v) + \lambda\mathbf{1} \ .$$

Proposition 1.22. *The above Bellman operators are monotone and additively homogeneous.*

Proposition 1.23. *Any monotone additively homogeneous operator $\mathcal{B} : \mathbb{R}^{\mathcal{E}} \to \mathbb{R}^{S}$ is nonexpansif, that is Liptschitz continuous with Lipschitz constant* 1*, for the sup-norm ($\|v\|_\infty = \sup\{|v(x)| \mid x \in \mathcal{E}\}$):*

$$\|\mathcal{B}(v) - \mathcal{B}(w)\|_\infty \leq \|v - w\|_\infty \quad \forall v, w \in \mathbb{R}^{\mathcal{E}} \ .$$

Proof. For all $v, w \in \mathbb{R}^{\mathcal{E}}$, we have $v(x) - w(x) \leq \|v - w\|_\infty$, for all $x \in \mathcal{E}$, hence $v - w \leq \|v - w\|_\infty \mathbf{1}$.

So $v \leq \|v - w\|_\infty \mathbf{1} + w$. Using monotonicity of $\mathcal{B}$, and then additive homogeneity, we get

$$\mathcal{B}(v) \leq \mathcal{B}(\|v - w\|_\infty \mathbf{1} + w) \leq \|v - w\|_\infty \mathbf{1} + \mathcal{B}(w) \ .$$

Hence $\mathcal{B}(v) - \mathcal{B}(w) \leq \|v - w\|_\infty \mathbf{1}$, that is $[\mathcal{B}(v) - \mathcal{B}(w)](x) \leq \|v - w\|_\infty$, for all $x \in \mathcal{E}$.

By symmetry, we also get $[\mathcal{B}(w) - \mathcal{B}(v)](x) \leq \|w - v\|_\infty$, and deduce $\|\mathcal{B}(v) - \mathcal{B}(w)\|_\infty \leq \|v - w\|_\infty$. □

Remark 1.24. Another proof of the nonexpansivity of the operator $\mathcal{B}$ of a deterministic control problem is as follows. Such an operator acts on $\mathbb{R}^{\mathcal{E}}$ and satisfies

$$\begin{aligned}[\mathcal{B}(v)](x) &= \sup\{r(x, u) + v(f(x, u)) \mid u \in \mathcal{C}(x)\} \\ &= \sup\{G(x, y) + v(y) \mid y \in \mathcal{E}\} \ ,\end{aligned}$$

for all $v \in \mathbb{R}^{\mathcal{E}}$ and $x \in \mathcal{E}$, for some control sets $\mathcal{C}(x)$ and maps r, f and G. Given $v, w \in \mathbb{R}^{\mathcal{E}}$ and $x \in \mathcal{E}$, choose $y \in \mathcal{E}$ which is optimal in the expression of $[\mathcal{B}(v)](x)$, that is such that $[\mathcal{B}(v)](x) = G(x, y) + v(y)$. Since $[\mathcal{B}(w)](x) \geq G(x, y) + w(y)$, we obtain

$$[\mathcal{B}(v)](x) - [\mathcal{B}(w)](x) \leq v(y) - w(y) \leq \|v - w\|_\infty \ .$$

By symmetry, we also get $[\mathcal{B}(w)](x) - [\mathcal{B}(v)](x) \leq \|v - w\|_\infty$, and taking the maximum over $x \in \mathcal{E}$, we deduce $\|\mathcal{B}(v) - \mathcal{B}(w)\|_\infty \leq \|v - w\|_\infty$.

1.5 Infinite horizon problems

Assume now that the horizon T is infinite, that is the functional J is replaced by:

$$J(X; U) := \sum_{k=0}^{\infty} r_k(X_k, U_k) \tag{1.6}$$

and that the infinite sum is well defined in $\mathbb{R} \cup \{-\infty\}$ for all sequences X_k and U_k satisfying (1.2b). Then, one may try to compute again the maximum (or supremum) v of J, as in finite horizon problems.

This is the case when

(A1) $\mathcal{C}_k$ and f_k do not depend on k (then we omit k in the notations), $r_k(x,u) = \alpha^k r(x,u)$ for all $k \geq 0$, for some function r bounded from above and some constant $\alpha > 0$;

and one of the following assumptions hold:

(A2) $\alpha < 1$;

(A3) $\alpha = 1$ and $r(x,u) \leq 0$ for all $x \in \mathcal{E}$ and $u \in \mathcal{C}$, and for all $x_0 \in \mathcal{E}$, there exists a sequence (X_k, U_k) satisfying (1.2b), such that $r(X_k, U_k) = 0$ for k large enough.

(A4) $\alpha = 1$, if G is constructed as in Section 1.4.1, then any circuit of the graph of G has a total weight ≤ 0, and for all $x_0 \in \mathcal{E}$, there exists an infinite path $x_0, x_1, \ldots$ in the graph of G, such that $G(x_k, x_{k+1}) = 0$ for k large enough.

Definition 1.25. The parameter α is called the *discount factor*. We say that the infinite horizon optimal control problem is *discounted* if $\alpha < 1$, and that it is *undiscounted* when $\alpha = 1$.

Definition 1.26. We define strategies, feedback policies and open-loop controls of an infinite horizon problem as for finite horizon problems, but with $T = \infty$.

We say that a feedback policy $\pi_* = (\pi_k)_{k\geq 0}$ of an infinite horizon control problem is *stationary* if $\pi_k = \pi_0$ for all $k \geq 0$.

Moreover we will sometimes use the same notation for π_* and π_0.

Theorem 1.27 (Stationnary deterministic dynamic programming)**.** *Assume that $\mathcal{E}$ is a finite set, and that* (A1) *holds together with one of the assumptions* (A2)*,* (A3) *or* (A4)*. Then the value function v of the problem*

$$v(x_0) = \sup\{J(X;U) \mid (X,U) \ \textit{satisfies (1.2b)}\ \} \tag{1.7}$$

with J as in (1.6)*, satisfies the equation:*

$$v(x) = \sup\{r(x,u) + \alpha v(f(x,u)) \mid u \in \mathcal{C}(x)\} \quad \forall x \in \mathcal{E}\ . \tag{1.8}$$

For all $x_0 \in \mathcal{E}$, the value $v(x_0)$ coincides with the optimum of J over all strategies or over all feedback policies, or over all open-loop controls. When $\alpha < 1$, the solution of (1.8) *is unique.*

Moreover, assume that the maximum of (1.8) *is attained for an action $u \in \mathcal{C}(x)$ and* let us denote by $\pi(x)$ this action*, then the stationary feedback policy $\pi_* = (\pi_k)_{k\geq 0}$, with $\pi_k = \pi$ for all $k \geq 0$, is an optimal strategy of the problem, and the dynamics $X_{k+1} = f(X_k, \pi(X_k))$ with $U_k = \pi(X_k)$ furnishes an optimal solution (X,U) of the infinite horizon control problem.*

To show this result, we shall use the Bellman operator.

Definition 1.28. Let $\mathcal{B}_\alpha : \mathbb{R}^{\mathcal{E}} \to \mathbb{R}^{\mathcal{E}}$ be the map such that for all $v \in \mathbb{R}^{\mathcal{E}}$, and $x \in \mathcal{E}$, we have

$$\begin{aligned}[\mathcal{B}_\alpha(v)](x) &= \sup\{r(x,u) + \alpha v(f(x,u)) \mid u \in \mathcal{C}(x)\} \\ &= \sup\{G(x,y) + \alpha v(y) \mid y \in \mathcal{E}\}\ ,\end{aligned}$$

where

$$G(x,y) = \sup\{r(x,u) \mid u \in \mathcal{C}(x),\ f(x,u) = y\} \in \mathbb{R} \cup \{-\infty\}\ .$$

The map $\mathcal{B}_\alpha$ is called the *Bellman operator* of the discounted infinite horizon optimal control problem.

Note that since r is bounded from above, $G(x,y)$ exists in $\mathbb{R}\cup\{-\infty\}$. Then, if $\mathcal{E}$ a finite set, we get that the values of $G(x,y)$ which are finite (that is $\neq -\infty$) are bounded from below. Then, one may have assumed from the begining that r is bounded from below and above.

The dynamic programming equation can be rewritten in functional form as the *fixed point equation* of the Bellman operator $\mathcal{B}_\alpha$:

$$v = \mathcal{B}_\alpha(v) \ .$$

Fact 1.29. The undiscounted Bellman operator $\mathcal{B}_1$ is monotone and additively homogenous.

Corollary 1.30. *The discounted Bellman operator $\mathcal{B}_\alpha$ is Lipschitz continuous for the sup-norm with Lipschitz constant α, thus it is α-contracting when $\alpha<1$.*

Proof. From Proposition 1.23, $\mathcal{B}_1$ is nonexpansive for the sup-norm. We have $\mathcal{B}_\alpha(v)=\mathcal{B}_1(\alpha v)$, so $\|\mathcal{B}_\alpha(v)-\mathcal{B}_\alpha(w)\|_\infty = \|\mathcal{B}_1(\alpha v)-\mathcal{B}_1(\alpha w)\|_\infty \le \|\alpha v-\alpha w\|_\infty = \alpha\|v-w\|_\infty$. □

Corollary 1.31. *When $\mathcal{E}$ is finite and $\alpha<1$, the operator $\mathcal{B}_\alpha$ admits a unique fixed point v^*. Moreover, for any initial point $v_0\in\mathbb{R}^{\mathcal{E}}$, the sequence $v_{n+1}=\mathcal{B}_\alpha(v_n)$ converges towards v^*:*

$$\|v_n - v^*\|_\infty \le \alpha^n \|v_0 - v^*\|_\infty \ .$$

Proof. This follows from the fixed point theorem since $\mathbb{R}^{\mathcal{E}}$ is a Banach space and $\mathcal{B}_\alpha$ is contracting. □

Proof of Theorem 1.27 when $\alpha<1$. Assume that $\mathcal{E}$ is a finite set and that $\alpha<1$. Let v^* be the unique solution of the Bellman equation $v=\mathcal{B}_\alpha(v)$, by Corollary 1.31. Let $v^{(N)}$ be the value function of the finite horizon problem:

$$v^{(N)}(x) = \max_{X,U}\{J^{(N)}(X;U) \mid (X_k,Y_k) \text{ satisfying (1.2b)}\}$$

with

$$J^{(N)}(X;U) := \left(\sum_{k=0}^{N-1}\alpha^k r(X_k,U_k)\right) + 0 \ .$$

From Theorem 1.12, $v^{(N)}=v_0^{(N)}$ with $v_k^{(N)}$ solution of the dynamic programming equation:

$$\begin{aligned} &v_N^{(N)}(x) = 0 \quad \forall x\in\mathcal{E} \ , \\ &v_k^{(N)}(x) = \sup\{\alpha^k r(x,u) + v_{k+1}^{(N)}(f(x,u)) \mid u\in\mathcal{C}(x)\} \quad \forall x\in\mathcal{E},\ k\le N-1. \end{aligned}$$

This can be rewritten as $v_k^{(N)} = \alpha^k \mathcal{B}_\alpha(v_{k+1}^{(N)}/\alpha^{k+1})$. Hence, $v^{(N)}=\mathcal{B}_\alpha^N(0) := \mathcal{B}_\alpha\circ\cdots\circ\mathcal{B}_\alpha(0)$ (where the composition is done N times). Therefore, $\lim_{N\to\infty} v^{(N)} = v^*$ where the limit is uniform in $\mathcal{E}$ (limit for the sup-norm of $\mathbb{R}^{\mathcal{E}}$).

Let C be a bound of the finite values of $|G(x,y)|$, with G as in Definition 1.28. Then, for any infinite sequences X and U, we have

$$|J^{(N)}(X;U) - J(X;U)| \le \sum_{k=N}^{\infty}\alpha^k C = \alpha^N \frac{C}{1-\alpha} \ .$$

Using the definition of the value function v of the infinite horizon problem and that of $v^{(N)}$, as the supremum of $J(X;U)$ and $J^{(N)}(X;U)$ respectively, we deduce

$$\|v - v^{(N)}\|_\infty \leq \alpha^N \frac{C}{1-\alpha} \ ,$$

so $\lim_{N\to\infty} v^{(N)} = v$, which implies that $v = v^*$.

The proof of optimality is similar to the finite horizon case. □

Proof of Theorem 1.27 when $\alpha = 1$. Under Assumption (A3), we get that $v \leq 0$ and $v(x) \in \mathbb{R}$ for all $x \in \mathcal{E}$. Moreover, $v(x) \leq v^{(N)}(x) \leq v^{(N-1)}(x)$ for all $x \in \mathcal{E}$ and $N \geq 1$. This implies that $v^{(N)}$ has a limit v^* which satisfies $v \leq v^*$. Since $v^{(N)} = \mathcal{B}_1(v^{(N-1)})$ for all $N \geq 1$, and $\mathcal{B}_1$ is continuous, we get that $v^* = \mathcal{B}_1(v^*)$.

It remains to prove that $v = v^*$. We shall use the finiteness of $\mathcal{E}$. Let $\delta > 0$ be a lower bound of $-G(x,y)$ over all (x,y) such that $G(x,y) < 0$ and let n be the cardinality of $\mathcal{E}$. Let $\varepsilon > 0$ be such that $\varepsilon < \delta$, and N such that $v^{(N)}(x) \leq v^*(x) + \varepsilon/3$. We get that $v^{(N)}(x) \leq v^{(N+n)}(x) + \varepsilon/3$ and if (X,U) is $\varepsilon/3$-optimal for $v^{(N+n)}$, we deduce that $\sum_{k=0}^{N-1} G(X_k, X_{k+1}) \leq v^{(N)}(x) \leq v^{(N+n)}(x) + \varepsilon/3 \leq \sum_{k=0}^{N+n-1} G(X_k, X_{k+1}) + 2 \times \varepsilon/3$ so, for all $k = N, \dots N+n-1$, we have $-G(X_k, X_{k+1}) \leq 2 \times \varepsilon/3$. This implies that all these (X_k, X_{k+1}) are such that $G(X_k, X_{k+1}) = 0$, and since the cardinality of $\mathcal{E}$ is equal to n, two elements of the sequence $(X_N, \dots, X_{N+n})$ are equal, which means that there is a cycle $(X_\ell, \dots, X_{\ell'} = X_\ell)$. Consider the infinite sequence obtained by concatening $(X_0, \dots, X_\ell)$ with an infinite number of the cycle $(X_\ell, \dots X_{\ell'})$. We obtain that $J(X;U) = \sum_{k=0}^{\ell-1} G(X_k, X_{k+1}) = \sum_{k=0}^{N+n-1} G(X_k, X_{k+1}) \geq v^{(N+n)}(x) - \varepsilon/3 \geq v^* - \varepsilon/3$. Since $v(x) \geq J(X;U)$ we deduce that $v(x) \geq v^*(x) - \varepsilon/3$. Since this holds for all $\varepsilon > 0$ (with $\varepsilon < \delta$), we get that $v(x) \geq v^*(x)$, and so the equality. □

Definition 1.32. The algorithm constructing the sequence $v_{n+1} = \mathcal{B}_\alpha(v_n)$ is called *value iterations*.

In practice one uses rather a variant similar to Gauss-Seidel algorithm (wrt to Jacobi) for the solution of linear systems, in order to avoid useless storage. The resulting algorithm is called *Ford-Bellman algorithm*. It depends on some ordering on $\mathcal{E}$. When $\mathcal{E} = \{1, \dots, N\}$, it is as follows:

$$\begin{aligned}
&v_{k,0} = v_k, \\
&\text{for } j = 1, \dots, N,\ v_{k,j}(j) = [\mathcal{B}_\alpha(v_{k,j-1})](j), \quad v_{k,j}(\ell) = v_{k,j-1}(\ell)\ \forall \ell \neq j, \\
&v_{k+1} = v_{k,N}.
\end{aligned}$$

When $\alpha = 1$, under suitable conditions, the value iterations converge in finite time $\leq N = \text{card}(\mathcal{E})$ to the solution. So with a computational time in $O(mN)$, where $m = \text{card}(\{(x,y) \in \mathcal{E} \times \mathcal{E} \mid G(x,y) \neq -\infty\})$.

When in addition (A3) holds (in particular $r \leq 0$), the problem is equivalent to a shortest path problem with weight G and no constraints in the length of paths.

The fixed point equation can be solved using Dijkstra algorithm which is equivalent to one full step of Ford-Bellman algorithm with an appropriate ordering of states. The computational time is then in $\mathcal{O}(m + N \log N)$ (with the implementation of Fredman and Tarjan).

Exercise 1.5.1 (Hierarchical shortest path problem). Alice and Bob are in holidays in Venezia with the little Charlie (1 year). Venizia is composed of islets connected with bridges with several stairs, which are thus difficult to cross with the heavy stroller of Charlie.

Assume that the map of Venezia is approximated by a graph in which nodes correspond to landmarks and arcs correspond to streets and bridges, and that we know the travel time and number of stairs to cross between nodes. Alice et Bob want to find the paths between two points x_i and x_f of the graph which minimize first the number of stairs and among all paths minimizing the number of stairs, they want to choose the ones which minimize also the travel time. Modelize this problem as a deterministic control problem.

1.6 Max-plus or Tropical algebra

The equations in (1.3b) can be seen as linear over the following semifield.

Consider the set $\mathbb{R} \cup \{-\infty\}$ of real numbers extended by $-\infty$, endowed with the maximization as an addition and the usual addition as a multiplication: $a \oplus b = \max(a, b)$ and $a \otimes b = a + b$ for all $a, b \in \mathbb{R} \cup \{-\infty\}$.

The addition is commutative, associative and has the zero element $-\infty$, which is absorbing for the multiplication, the multiplication is commutative, associative, has the unit (neutral) element 0, and it distributes over the addition $\oplus$.

The addition is idempotent, meaning that $a \oplus a = a$ for all a. Therefore opposites to non zero elements do not exist.

Inverses (for the multiplication) exist for all non zero elements. So we obtain a semifield called the *max-plus algebra* or the *tropical algebra*, that is often denoted $\mathbb{R}_{\max}$.

Then the dynamic programming equation of deterministic control with finite horizon (1.3b) can be rewritten as

$$v_k(x) = \bigoplus_{u \in \mathcal{C}_k(x)} r_k(x,u) \otimes v_{k+1}(f_k(x,u)) \quad \forall x \in \mathcal{E}, \text{ and } k = T-1, \dots, 0 \ , \tag{1.9}$$

which is a linear equation over $\mathbb{R}_{\max}$. In particular, denoting

$$M_{xy}^{(k)} = \sup\{r_k(x,u) \mid u \in \mathcal{C}_k(x),\ f_k(x,u) = y\}$$

for all $x, y \in \mathcal{E}$ (this was G_k above), then (1.9) can be rewritten as

$$v_k(x) = \bigoplus_{y \in \mathcal{E}} M_{xy}^{(k)} \otimes v_{k+1}(y) \quad \forall x \in \mathcal{E}, \text{ and } k = T-1, \dots, 0 \ ,$$

that is the vector v_k is the product of the tropical matrix $M^{(k)}$ with entries $M_{xy}^{(k)}, x, y \in \mathcal{E}$ by the vector v_{k+1}. Hence, v_0 is obtained by applying the product of the T matrices $M^{(0)}, \dots, M^{(T-1)}$ to the vector $v_T = \varphi$.

In this way, the Bellman equation can be seen as a Kolmogorov equation associated to a Markov chain, as in Chapter 2.

In some situations, the analogy with usual numerical linear algebra may suggest some algorithms. These algorithms are often called tropical numerical methods.

1.7 Solutions of Exercises

Exercise 1.3.1. The example can be solved by applying (1.5) with $N = 3$, $m_1 = 4, m_2 = 3, m_3 = 2$, $w_1 = 5, w_2 = 4, w_3 = 3$. These equations reduce to:

$$\begin{aligned} v_3(x) &= 0 \quad \text{for } x \in \{0, \dots, 10\} \\ v_2(x) &= \max(v_3(x), 2 + v_3(x-3)) \\ v_1(x) &= \max(v_2(x), 3 + v_2(x-4)) \\ v_0(x) &= \max(v_1(x), 4 + v_1(x-5)) \end{aligned}$$

with the extension of v_k to $-\infty$ for $x < 0$. This gives the following table of values of the problem:

x	0	1	2	3	4	5	6	7	8	9	10
v_3	0	0	0	0	0	0	0	0	0	0	0
v_2	0	0	0	2	2	2	2	2	2	2	2
v_1	0	0	0	2	3	3	3	5	5	5	5
v_0	0	0	0	2	3	4	4	5	6	7	7

This table determines the optimal policy at each step $k = 0, 1, 2$: $\pi_k(x) = 0$ if $v_k(x) = v_{k+1}(x)$ and $\pi_k(x) = 1$ otherwise.

The value of the knapsack problem is equal to $v_0(10) = 7$. It is obtained by starting with $X_0 = 10$ and taking the controls $u_k = U_{k-1} = \pi_{k-1}(X_{k-1})$, and $X_k = X_{k-1} - w_{k-1}U_{k-1}$. In view of the above table, we have $v_0(10) \neq v_1(10)$, so $u_1 = U_0 = 1$ and $X_1 = 10 - 5 = 5$. For $X_1 = 5$, we have $v_1(5) = 3 \neq v_2(5)$, so $u_2 = U_1 = 1$ and $X_2 = 5 - 4 = 1$. For $X_2 = 1$, we have $v_2(1) = 0 = v_3(1)$, so $u_3 = U_2 = 0$ and $X_3 = 1$.

Exercise 1.5.1.

Chapter 2

Markov chains and Kolmogorov equations

2.1 Introduction and Notations

We shall consider here Markov chains over a finite (or countable) state space $\mathcal{E}$. These are the random version of the dynamical system $X_{n+1} = f_n(X_n)$.

We shall also consider functionals similar to the ones optimized in the optimal control problems of Chapter 1. and prove a "linear version" of Bellman dynamic programming equation: the Kolmogorov equation.

Bellow are some general notations used in all the course.

- The state space $\mathcal{E}$ is assumed to be finite or possibly countable, that is *discrete*. Let $N = \mathrm{card}(\mathcal{E}) \in \mathbb{N} \cup \{\infty\}$.
- The elements of $\mathcal{E}$ are (linearly) ordered, one can identify $\mathcal{E}$ with $\{1, \dots, N\}$ (or $\mathbb{N}$ if $N = \infty$), and any element of $\mathbb{R}^{\mathcal{E}}$, that is any function $\mathcal{E} \to \mathbb{R}$, to a *column vector* in $\mathbb{R}^N$.
- We shall use this identification, without fixing any order on $\mathcal{E}$.
- More generally, we shall speak about matrices over $\mathcal{E}$: an element M of $\mathbb{R}^{\mathcal{E}\times\mathcal{E}}$ is a matrix over $\mathcal{E}$, and its entries are denoted $(M_{xy})_{x,y\in\mathcal{E}}$.
- $\mathcal{E}$ beeing at most countable, we shall endow $\mathcal{E}$ with the σ-algebra $\mathcal{P}(\mathcal{E})$ of all subsets of $\mathcal{E}$.
- A probability law p over $\mathcal{E}$ will be identified to a row vector, although we will also write $p \in \mathbb{R}^{\mathcal{E}}$, more precisely this is an element of the *simplex*:

$$\Delta_{\mathcal{E}} = \{p \in \mathbb{R}^{\mathcal{E}} \mid p_x \geq 0 \ \forall x \in \mathcal{E},\ p\mathbf{1} = \sum_{x\in\mathcal{E}} p_x = 1\} \ .$$

- The *Dirac measure* over $\mathcal{E}$ in state $x \in \mathcal{E}$ will be denoted δ_x: its entries are 1 in x and 0 elsewhere.
- A matrix $M \in \mathbb{R}^{\mathcal{E}\times\mathcal{E}}$ is a *Markov (or a stochastic) matrix* if its entries are all nonnegative ($M_{xy} \geq 0$ for all $x, y \in \mathcal{E}$) and $M\mathbf{1} = \mathbf{1}$ ($\sum_{y\in\mathcal{E}} M_{xy} = 1$ for all $x \in \mathcal{E}$).

- Given a probability space $(\Omega, \mathfrak{A}, P)$, a random variable taking its values in $\mathcal{E}$ is by definition a measurable function from $(\Omega, \mathfrak{A})$ to $(\mathcal{E}, \mathcal{P}(\mathcal{E}))$.
- Given a filtration $(\mathcal{F}_n)_{n\in\mathbb{N}}$ on $(\Omega, \mathfrak{A})$, that is a nondecreasing sequence of σ-algebras $\mathcal{F}_n \subset \mathfrak{A}$, we say that a sequence $(X_n)_{n\geq 0}$ of random variables (also called a *random process*) taking its values in $\mathcal{E}$ is adapted to the filtration if for all $n \geq 0$, X_n is measurable from $(\Omega, \mathcal{F}_n)$ to $(\mathcal{E}, \mathcal{P}(\mathcal{E}))$.

2.2 Markov property

Definition 2.1. A sequence $(X_n)_{n\in\mathbb{N}}$ of random variables over $(\Omega, \mathfrak{A}, P)$, taking its values in $\mathcal{E}$, is a *Markov chain* (or a discrete time Markov process) if it satisfies:

$$P(X_{n+1} = x_{n+1} \mid X_0 = x_0, \ldots, X_n = x_n) = P(X_{n+1} = x_{n+1} \mid X_n = x_n) \tag{2.1}$$

for all $n \geq 1$, $x_0, \ldots, x_{n+1} \in \mathcal{E}$.

The probability measure $p^{(0)}$ and the Markov matrices $M^{(n)}$ over $\mathcal{E}$ defined by:

$$\begin{aligned} p_x^{(0)} &= P(X_0 = x) \\ M_{xy}^{(n)} &= P(X_{n+1} = y \mid X_n = x) \end{aligned}$$

for all $x, y \in \mathcal{E}$ are respectively called the *initial law* and the *transition matrix* at time n of the Markov chain $(X_n)_{n\in\mathbb{N}}$.

The Markov chain is *stationary* if the transition matrices $M^{(n)}$ do not depend on n. If $p^{(0)} = \delta_{x_0}$, we say that x_0 is the *initial state* of the Markov chain.

Example 2.2 (Random walk)**.** A particule or a drunk man is walking on a line: at each unit of time, he is going forward with probability $p \in [0, 1]$, and backward with probability $1 - p$. If he stops at boundaries of $\mathcal{E} = \{0, \ldots, N\}$, then the position X_n at time n defines a Markov chain X_n over $\mathcal{E}$ such that

$$P(X_{n+1} = x + 1 \mid X_n = x) = 1 - P(X_{n+1} = x - 1 \mid X_n = x) = p,$$

when $x \neq 0, N$.

The transition matrix is given by:

$$M = \begin{bmatrix} 1 & 0 & 0 & 0 & \cdots & 0 \\ 1-p & 0 & p & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & 1-p & 0 & p \\ 0 & \cdots & 0 & 0 & 0 & 1 \end{bmatrix}$$

Proposition 2.3. *Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of random variables over $(\Omega, \mathfrak{A}, P)$, taking its values in $\mathcal{E}$, with initial law $p^{(0)}$. Then the following are equivalent:*

1. *$(X_n)_{n\in\mathbb{N}}$ is a Markov chain with transition matrices $M^{(n)}$ at time $n \in \mathbb{N}$;*

2. $P(X_0 = x_0, \dots, X_n = x_n) = p^{(0)}_{x_0} M^{(0)}_{x_0 x_1} \cdots M^{(n-1)}_{x_{n-1} x_n}$, *for all* $n \geq 1$, $x_0, \dots, x_n \in \mathcal{E}$;

3. $P(X_{n+1} = x_{n+1} \mid X_0 = x_0, \dots, X_n = x_n) = M^{(n)}_{x_n x_{n+1}}$, *for all* $n \geq 0$, $x_0, \dots, x_{n+1} \in \mathcal{E}$;

Proof. Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of random variables over $(\Omega, \mathfrak{A}, P)$, taking its values in $\mathcal{E}$, with initial law $p^{(0)}$.

1. $\Rightarrow$ 2. Assume $(X_n)_{n\in\mathbb{N}}$ is a Markov chain with transition matrices $M^{(n)}$ at time $n \in \mathbb{N}$, then it satisfies:

$$\begin{aligned} & P(X_0 = x_0, \dots, X_n = x_n) \\ &= P(X_0 = x_0, \dots, X_{n-1} = x_{n-1}) P(X_n = x_n \mid X_0 = x_0, \dots, X_{n-1} = x_{n-1}) \\ &= P(X_0 = x_0, \dots, X_{n-1} = x_{n-1}) M^{(n-1)}_{x_{n-1} x_n} \ , \end{aligned}$$

from which one deduce 2. by induction.

2. $\Rightarrow$ 3. Assume that (X_n) satisfies 2. Then,

$$\begin{aligned} & P(X_{n+1} = x_{n+1} \mid X_0 = x_0, \dots, X_n = x_n) \\ &= \frac{P(X_0 = x_0, \dots, X_{n+1} = x_{n+1})}{P(X_0 = x_0, \dots, X_n = x_n)} \\ &= \frac{p^{(0)}_{x_0} M^{(0)}_{x_0 x_1} \cdots M^{(n)}_{x_n x_{n+1}}}{p^{(0)}_{x_0} M^{(0)}_{x_0 x_1} \cdots M^{(n-1)}_{x_{n-1} x_n}} \\ &= M^{(n)}_{x_n x_{n+1}} \ , \end{aligned}$$

that is 3.

3. $\Rightarrow$ 1. Assume now that (X_n) satisfies 3., or equivalently assume that $P(X_{n+1} = x_{n+1} \mid X_0 = x_0, \dots, X_n = x_n)$ does not depend on $x_0, \dots, x_{n-1}$. Let us deduce that it is also equal to $P(X_{n+1} = x_{n+1} \mid X_n = x_n)$, that is the Markov property. Indeed,

$$\begin{aligned} & P(X_n = x_n, X_{n+1} = x_{n+1}) \\ &= \sum_{x_0, \dots, x_{n-1} \in \mathcal{E}} P(X_0 = x_0, \dots, X_n = x_n, X_{n+1} = x_{n+1}) \\ &= \sum_{x_0, \dots, x_{n-1} \in \mathcal{E}} \left(P(X_0 = x_0, \dots, X_n = x_n) P(X_{n+1} = x_{n+1} \mid X_0 = x_0, \dots, X_n = x_n) \right) \\ &= \left(\sum_{x_0, \dots, x_{n-1} \in \mathcal{E}} P(X_0 = x_0, \dots, X_n = x_n) \right) M^{(n)}_{x_n x_{n+1}} \\ &= P(X_n = x_n) P(X_{n+1} = x_{n+1} \mid X_0 = x_0, \dots, X_n = x_n) \ . \end{aligned}$$

□

Corollary 2.4. *Let* $(X_n)_{n\in\mathbb{N}}$ *be a sequence of random variables over* $(\Omega, \mathfrak{A}, P)$*, taking its values in* $\mathcal{E}$*. If, for all* $n \geq 0$, $x_0, \dots, x_{n+1} \in \mathcal{E}$, $P(X_{n+1} = x_{n+1} \mid X_0 = x_0, \dots, X_n = x_n)$ *does not depend on* $x_0, \dots, x_{n-1}$*, that is is a function of* n, x_n, x_{n+1} *only, then* $(X_n)_{n\in\mathbb{N}}$ *is a Markov chain.*

Theorem 2.5. *Let $p^{(0)}$ be a probabilty over $\mathcal{E}$ and $M^{(n)}$ be Markov matrices over $\mathcal{E}$, for all $n \in \mathbb{N}$. Then, there exists a probability space $(\Omega, \mathfrak{A}, P)$ and a Markov chain $(X_n)_{n\in\mathbb{N}}$ on $(\Omega, \mathfrak{A}, P)$ taking its values in $\mathcal{E}$, with initial law $p^{(0)}$ and transition matrices $M^{(n)}$.*

Sketch of proof. We know that necessarily, the law of X_n is given by the formula in 2 of Proposition 2.3. This allows to compute the probability of all the cylinders $A_0 \times \cdots \times A_n \times \mathcal{E}^{\mathbb{N}}$ of $\mathcal{E}^{\mathbb{N}}$. Therefore, it is sufficient to consider the canonical probability space $\Omega = \mathcal{E}^{\mathbb{N}}$, with $\mathfrak{A}$ the σ-algebra generated by finite cylinders, and P the probability on $(\Omega, \mathfrak{A})$ already given on cylinders. Such a probability exists and is unique by Kolmogorov extension theorem. □

2.3 Elementary Properties and representations

Proposition 2.6 (Fokker-Plank equation)**.** *Let $(X_n)_{n\in\mathbb{N}}$ be a Markov chain over $(\Omega, \mathfrak{A}, P)$, taking its values in $\mathcal{E}$, with initial law $p^{(0)}$ and Markov transition matrices $M^{(n)}$ at time $n \in \mathbb{N}$. Then, the law $p^{(n)}$ of the random variable X_n satisfies the* Fokker-Plank *recurrence equation:*

$$p^{(n+1)} = p^{(n)} M^{(n)} \ .$$

Proof. Using the definition of $p^{(n)}$, $M^{(n)}$ and of conditional probabilities, we get, for all $x_{n+1} \in \mathcal{E}$,

$$\begin{aligned} p^{(n+1)}_{x_{n+1}} &= P(X_{n+1} = x_{n+1}) \\ &= \sum_{x_n \in \mathcal{E}} P(X_n = x_n, X_{n+1} = x_{n+1}) \\ &= \sum_{x_n \in \mathcal{E}} P(X_{n+1} = x_{n+1} \mid X_n = x_n) P(X_n = x_n) \\ &= \sum_{x_n \in \mathcal{E}} M^{(n)}_{x_n x_{n+1}} p^{(n)}_{x_n} = (p^{(n)} M^{(n)})_{x_{n+1}} \ . \quad \square \end{aligned}$$

Remark 2.7. The proof of Fokker-Plank equation does not use the Markov property, but only the value of $P(X_{n+1} = x_{n+1} \mid X_n = x_n)$.

Proposition 2.8. *Let $(X_n)_{n\in\mathbb{N}}$ be a Markov chain over $(\Omega, \mathfrak{A}, P)$, taking its values in $\mathcal{E}$, with initial law $p^{(0)}$ and Markov transition matrices $M^{(n)}$ at time $n \in \mathbb{N}$. Then, the sequence $(X_{n+k})_{n\in\mathbb{N}}$ is a Markov chain with initial law $p^{(k)}$ (given by Fokker-Plank equation) and Markov transition matrices $M^{(n+k)}$ at time $n \in \mathbb{N}$.*

Proof. By Proposition 2.3, the sequence $(X_n)_{n\in\mathbb{N}}$ satisfies (2). Therefore, for $k, n \in \mathbb{N}$, we have

$$P(X_0 = x_0, \ldots, X_{n+k} = x_{n+k}) = p^{(0)}_{x_0} M^{(0)}_{x_0 x_1} \cdots M^{(k-1)}_{x_{k-1} x_k} \cdots M^{(n+k)}_{x_{n+k-1} x_{n+k}} \ .$$

Taking the sum for all $x_0, \ldots, x_{k-1} \in \mathcal{E}$, we get

$$P(X_k = x_k, \ldots, X_{n+k} = x_{n+k}) = \left(\sum_{x_0, \ldots, x_{k-1} \in \mathcal{E}} p^{(0)}_{x_0} M^{(0)}_{x_0 x_1} \cdots M^{(k-1)}_{x_{k-1} x_k} \right) M^{(k)}_{x_k x_{k+1}} \cdots M^{(n+k)}_{x_{n+k-1} x_{n+k}} \ . \tag{2.2}$$

For $n = 0$ this equation writes

$$P(X_k = x_k) = \sum_{x_0,\dots,x_{k-1}\in\mathcal{E}} p^{(0)}_{x_0} M^{(0)}_{x_0x_1} \cdots M^{(k-1)}_{x_{k-1}x_k} \ .$$

Together with (2.2) for any $n \in \mathbb{N}$, this gives

$$\begin{aligned} P(X_k = x_k, \dots, X_{n+k} = x_{n+k}) &= P(X_k = x_k) M^{(k)}_{x_kx_{k+1}} \cdots M^{(n+k)}_{x_{n+k-1}x_{n+k}} \\ &= p^{(k)}_{x_k} M^{(k)}_{x_kx_{k+1}} \cdots M^{(n+k)}_{x_{n+k-1}x_{n+k}} \ , \end{aligned}$$

where $p^{(k)}$ is the law of X_k. Therefore, using Proposition 2.3, we get that $(X_{n+k})_{n\in\mathbb{N}}$ is a Markov chain with initial law $p^{(k)}$ and transition matrix $M^{(n+k)}$ at time $n \in \mathbb{N}$. Moreover, from Proposition 2.6, $p^{(k)}$ satisfies Fokker-Plank equation. □

Proposition 2.9. *Let $(X_n)_{n\in\mathbb{N}}$ be a Markov chain over $(\Omega, \mathfrak{A}, P)$, taking its values in $\mathcal{E}$, with initial law $p^{(0)}$ and Markov transition matrices $M^{(n)}$ at time $n \in \mathbb{N}$. Then, for all $x_0, \dots, x_T \in \mathcal{E}$, $0 \le k \le T$, we have*

$$P(X_{k+1} = x_{k+1}, \dots, X_T = x_T \mid X_0 = x_0, \dots, X_k = x_k) = M^{(k)}_{x_kx_{k+1}} \cdots M^{(T-1)}_{x_{T-1}x_T} \ .$$

Proposition 2.10. *Let $(X_n)_{n\in\mathbb{N}}$ be a Markov chain over $(\Omega, \mathfrak{A}, P)$, taking its values in $\mathcal{E}$, with initial law $p^{(0)}$ and Markov transition matrices $M^{(n)}$ at time $n \in \mathbb{N}$. Then, for all $k \ge 1$, the sequence $(X_{nk})_{n\in\mathbb{N}}$ is a Markov chain with initial law $p^{(0)}$ and Markov transition matrices $M^{(nk)} \cdots M^{((n+1)k-1)}$ at time $n \in \mathbb{N}$.*

Proof. From Proposition 2.3, the sequence $(X_n)_{n\in\mathbb{N}}$ satisfies (2). Then, for all $k \ge 1$ and $n \in \mathbb{N}$, we have

$$P(X_0 = x_0, \dots, X_{nk} = x_{nk}) = p^{(0)}_{x_0} M^{(0)}_{x_0x_1} \cdots M^{(nk)}_{x_{nk-1}x_{nk}} \ .$$

Taking the sum over all $x_1, \dots, x_{k-1}, x_{k+1}, \dots, x_{2k-1}, \dots x_{(n-1)k+1}, \dots, x_{nk-1} \in \mathcal{E}$, we get:

$$\begin{aligned} &P(X_0 = x_0, \dots, X_{mk} = x_{mk}, \dots, X_{nk} = x_{nk}) \\ &= \sum_{x_1,\dots,x_{k-1},x_{k+1},\dots,x_{2k-1},\dots x_{(n-1)k+1},\dots,x_{nk-1}\in\mathcal{E}} p^{(0)}_{x_0} M^{(0)}_{x_0x_1} \cdots M^{(nk)}_{x_{nk-1}x_{nk}} \ . \end{aligned}$$

Denoting $M^{(k,n)} := M^{nk} \cdots M^{((n+1)k-1)}$, we obtain

$$P(X_0 = x_0, \dots, X_{mk} = x_{mk}, \dots, X_{nk} = x_{nk}) = p^{(0)}_{x_0} M^{(k,0)}_{x_0x_k} \cdots M^{(k,n)}_{x_{(n-1)k}x_{nk}} \ ,$$

which with Proposition 2.3 shows the result. □

Practical examples are often constructed in the following way.

Fact 2.11. Let $(\Omega, \mathfrak{A}, P)$ be a probability space. Let X_0 be a random variable taking its values in $\mathcal{E}$, $(W_n)_{n\ge0}$ be a sequence of independent random variables taking its values in a finite set $\mathcal{W}$, and independent of X_0, and for all $n \in \mathbb{N}$, let $f_n : \mathcal{E} \times \mathcal{W} \to \mathcal{E}$ be a map. Then, the sequence X_n defined recursively by:

$$X_{n+1} = f_n(X_n, W_n), \quad n \in \mathbb{N} \ ,$$

is a Markov chain taking its values in $\mathcal{E}$.

Proof. Indeed X_n is a deterministic function of $X_0, W_0, \ldots, W_{n-1}$, so is independent of W_n. Hence, $P(X_{n+1} = x_{n+1} \mid X_0 = x_0, \ldots, X_n = x_n) = P(f_n(X_n, W_n) = x_{n+1} \mid X_0 = x_0, \ldots, X_n = x_n) = P(f_n(x_n, W_n) = x_{n+1} \mid X_0 = x_0, \ldots, X_n = x_n) = P(f_n(x_n, W_n) = x_{n+1})$ only depends on x_n and x_{n+1}. From Proposition 2.3, X_n is a Markov chain with transition matrix $M^{(n)}_{x_n x_{n+1}} = P(f_n(x_n, W_n) = x_{n+1})$ □

Proposition 2.12. *Conversely, let $p^{0)}$ and $M^{(n)}$ be the initial law and transition Markov matrices of a Markov chain taking its values in $\mathcal{E}$. There exists a probability space $(\Omega, \mathfrak{A}, P)$, a set $\mathcal{W}$, a sequence $(f_n)_{n\geq 0}$ of maps $f_n : \mathcal{E} \times \mathcal{W} \to \mathcal{E}$, and, over $(\Omega, \mathfrak{A}, P)$, a random variable X_0 taking its values in $\mathcal{E}$ with law $p^{(0)}$, and a sequence $(W_n)_{n\geq 0}$ of independent random variables, taking their values in $\mathcal{W}$, and independent of X_0, such that the sequence $(X_n)_{n\geq 0}$ defined recursively by $X_{n+1} = f_n(X_n, W_n)$ is a Markov chain with transition Markov matrices $M^{(n)}$.*

Proof. Assume to simplify that $\mathcal{E}$ is a finite set. Let $\mathcal{W}$ be the set of maps from $\mathcal{E}$ to itself. For each $n \geq 0$, denote by q_n the probability law on $\mathcal{W}$ defined, for all $w \in \mathcal{W}$ ($w : \mathcal{E} \to \mathcal{E},\ x \mapsto w(x)$), by $q_n(w) = \prod_{x\in\mathcal{E}} M^{(n)}_{xw(x)}$. Let $X_0, W_0, \ldots, W_n, \ldots$ be independent random variables with values in $\mathcal{E}$, $\mathcal{W}$,..., $\mathcal{W}$, ..., respectively with laws $p^{(0)}$, q_0, ..., q_n, ..., respectively. Such a sequence can be constructed on $\Omega = \mathcal{E} \times \mathcal{W}^{\mathbb{N}}$, with $\mathfrak{A}$ the set of cylinders. Then, taking $f_n(x, w) = w(x)$, we get that $X_{n+1} = f_n(X_n, W_n)$ defines a Markov chain with transition Markov matrices $M^{(n)}$. □

Exercise 2.3.1. Let $(\Omega, \mathfrak{A}, P)$ be a probability space. Let X_0 be a random variable taking its values in a finite set $\mathcal{E}$, let $(W_n)_{n\geq 0}$ be a Markov chain taking its values in a finite set $\mathcal{W}$, with initial law $q^{(0)}$, transition matrix $M^{(n)}$ at time $n \geq 0$, and independent of X_0, and for all $n \in \mathbb{N}$, let $f_n : \mathcal{E} \times \mathcal{W} \to \mathcal{E}$ be a map. Consider the sequence X_n of random variables taking its values in $\mathcal{E}$, and defined recursively by:

$$X_{n+1} = f_n(X_n, W_n), \quad n \in \mathbb{N}\ .$$

Show that $((X_n, W_n))_{n\geq 0}$ is a Markov chain and compute its transition matrix.

2.4 The digraph of a stationary Markov chain

Definition 2.13. Let $M \in \mathbb{R}^{\mathcal{E}\times\mathcal{E}}$ be a matrix with nonnegative entries, in particular a Markov matrix. We associate to M a *digraph*, denoted $\mathcal{G}(M)$, with set of nodes $\mathcal{E}$ and set of arcs $\mathcal{A}$ such that $(x, y) \in \mathcal{E} \times \mathcal{E}$ is in $\mathcal{A}$ if and only if $M_{xy} > 0$.

We associate also the weight map $w : \mathcal{A} \to \mathbb{R},\ (x, y) \mapsto M_{xy}$.

Then, the weight of a path $p = (x_0, \ldots, x_n)$ of $\mathcal{G}(M)$ (i.e. such that $x_0, \ldots, x_n \in \mathcal{E}$ and $(x_0, x_1), \ldots, (x_{n+1}, x_n)$ are arcs in $\mathcal{G}(M)$), is defined as $w(p) = w((x_0, x_1)) \times \cdots \times w((x_{n-1}, x_n)) = M_{x_0x_1} \cdots M_{x_{n-1}x_n}$.

If X_n is a stationary Markov chain with transition matrix M, we have

$$P(X_1 = x_1, \ldots, X_n = x_n \mid X_0 = x_0) = \begin{cases} w(x_0, \ldots, x_n) \text{ if } (x_0, \ldots, x_n) \text{ is a path of } \mathcal{G}(M) \\ 0 \qquad\qquad\qquad\qquad \text{otherwise.} \end{cases}$$

Moreover,

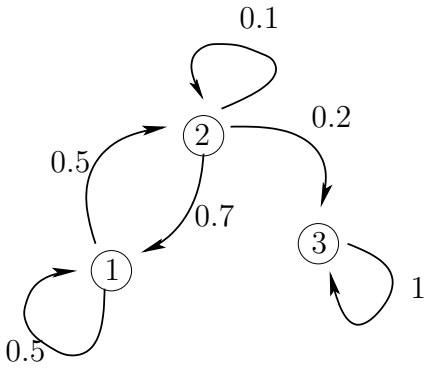
$$P(X_{n+k} = y \mid X_k = x) = (M^n)_{xy} = \sum_{\substack{p \text{ path of length } n \\ \text{in } \mathcal{G}(M), \text{ from } x \text{ to } y}} w(p),$$

where the length of a path is the number of its arcs. Indeed,

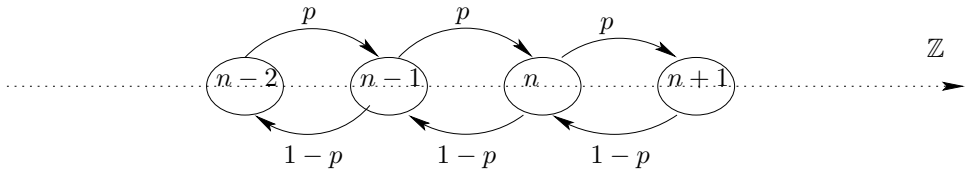
$$\begin{aligned} M^n_{xy} &= \sum_{(x_1,\dots,x_{n-1})\in\mathcal{E}^{n-1}} M_{xx_1}\cdots M_{x_{n-2}x_{n-1}}M_{x_{n-1}y} \\ &= \sum_{(x_1,\dots,x_{n-1})\in\mathcal{E}^{n-1}} P(X_{k+1}=x_1,\dots X_{k+n-1}=x_{n-1},\ X_{k+n}=y \mid X_k=x) \\ &= \sum_{\substack{(x_1,\dots,x_{n-1})\in\mathcal{E}^{n-1}\\ M_{xx_1}>0,\dots,M_{x_{n-1}y}>0}} w(x,x_1,\dots,x_{n-1},y)\ . \end{aligned}$$

Example 2.14. A Markov matrix and its associated digraph $\mathcal{G}(M)$ (with weights written on arcs):

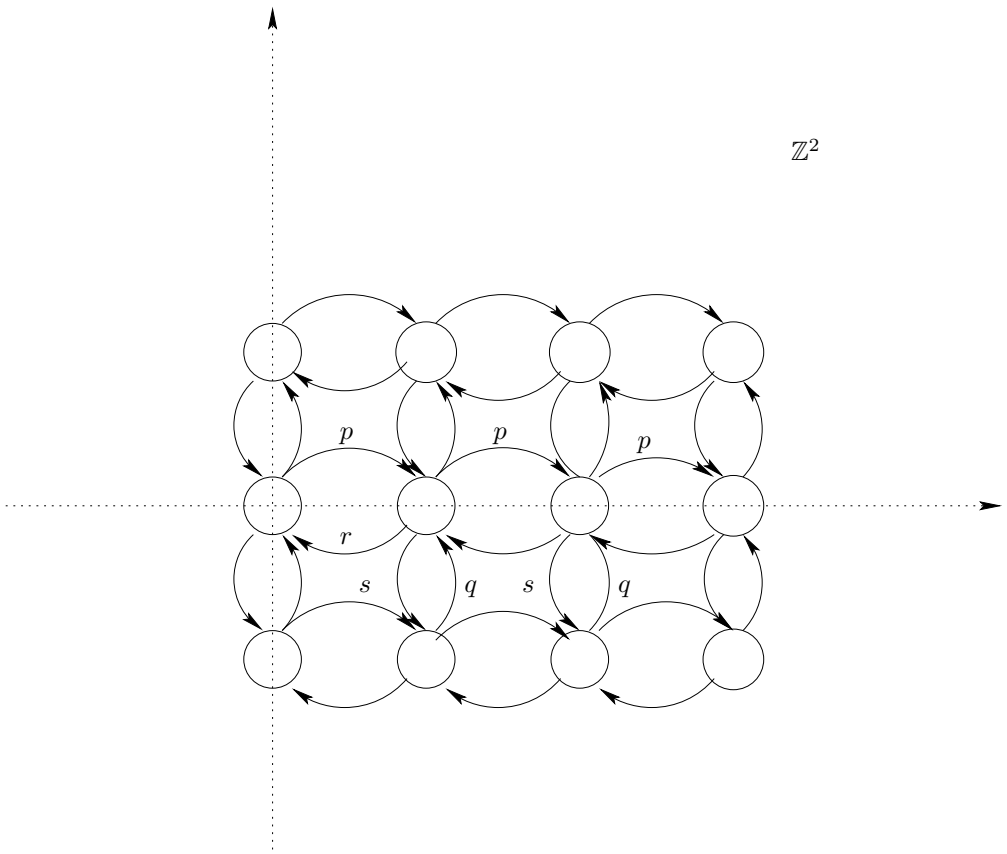
$$M=\begin{bmatrix} 0.5 & 0.5 & 0 \\ 0.7 & 0.1 & 0.2 \\ 0 & 0 & 1 \end{bmatrix}$$



Example 2.15 (Random walks)**.** The digraph associated to the random walk of Example 2.2 is as follows:



The digraph associated to a random walk on $\mathbb{Z}^2$ with probability $p,q,r,s>0$ to go respectively to right, up, left and down (with $p+q+r+s=1$) is shown in



2.5 Kolmogorov equation for finite horizon criteria

The following equation is the dual of Fokker-Plank equation (seen in Proposition 2.6):

Proposition 2.16 (Kolmogorov equation without instantaneous reward)**.** *Let $(X_n)_{n\in\mathbb{N}}$ be a Markov chain over $(\Omega, \mathfrak{A}, P)$, taking its values in $\mathcal{E}$, with Markov transition matrices $M^{(n)}$ at time $n \in \mathbb{N}$. Let $\varphi \in \mathbb{R}^{\mathcal{E}}$ and $T \in \mathbb{N}$, and denote:*

$$v_k(x) = \mathbb{E}\left[\varphi(X_T) \mid X_k = x\right] \ .$$

Then, the value v_k satisfies the following backward recurrence equation, called Kolmogorov equation *:*

$$v_k = M^{(k)} v_{k+1},\ 0 \leq k \leq T-1 \ ,$$

with final condition:

$$v_T = \varphi \ .$$

Proof. One way is to use the formula already proved for the probabilities of the tuple $(X_0, \dots, X_n)$:

$$\begin{aligned} v_k(x) &= \sum_{x_{k+1},\dots,x_T \in \mathcal{E}} (P(X_{k+1} = x_{k+1}, \dots, X_T = x_T \mid X_k = x)\varphi(x_T)) \\ &= \sum_{x_{k+1},\dots,x_T \in \mathcal{E}} \left(\frac{P(X_k = x, X_{k+1} = x_{k+1}, \dots, X_T = x_T)}{P(X_k = x)} \varphi(x_T) \right) \\ &= \sum_{x_{k+1},\dots,x_T \in \mathcal{E}} \left(M^{(k)}_{x x_{k+1}} \cdots M^{(T-1)}_{x_{T-1} x_T} \varphi(x_T) \right) \\ &= (M^{(k)} \cdots M^{(T-1)} \varphi)_x \ . \end{aligned}$$

Then $v_k = M^{(k)} \cdots M^{(T-1)}\varphi$, which implies $v_k = M^{(k)} v_{k+1}$ and $v_T = \varphi$.

Another way is to obtain Kolmogorov equation directely from conditional expectations:

$$v_k(x) = \mathbb{E}\left[\varphi(X_T) \mid X_k = x\right] = \mathbb{E}\left[\mathbb{E}\left[\varphi(X_T) \mid X_{k+1}\right] \mid X_k = x\right] \ .$$

$\mathbb{E}\left[\varphi(X_T) \mid X_{k+1}\right]$ is the projection of $\varphi(X_T)$ on the σ-algebra genererated by the random variable X_{k+1}. Since $\mathcal{E}$ is finite, we also get $\mathbb{E}\left[\varphi(X_T) \mid X_{k+1}\right] = h(X_{k+1})$, where h is the deterministic function defined by $h : \mathcal{E} \to \mathbb{R},\ x \mapsto \mathbb{E}\left[\varphi(X_T) \mid X_{k+1} = x\right]$, that is $h = v_{k+1}$. We deduce:

$$\begin{aligned} v_k(x) &= \mathbb{E}\left[v_{k+1}(X_{k+1}) \mid X_k = x\right] \\ &= \sum_{y \in \mathcal{E}} P(X_{k+1} = y \mid X_k = x) v_{k+1}(y) \\ &= \sum_{y \in \mathcal{E}} M^{(k)}_{xy} v_{k+1}(y) = (M^{(k)} v_{k+1})_x \ , \end{aligned}$$

so $v_k = M^{(k)} v_{k+1}$. □

Remark 2.17. Another way to prove Proposition 2.16 is to show that the sequence

$$\mathcal{M}_n = v_n(X_n), \quad n \geq 0,$$

is a Martingale for the filtration $(\mathcal{F}_n)_{n \geq 0}$ associated to the Markov chain $(X_n)_{n \geq 0}$, that is

$$\mathbb{E}\left[\mathcal{M}_{n+1} \mid \mathcal{F}_n\right] = \mathcal{M}_n \ .$$

Indeed, using the definition of v_n, we get, for $n \geq 0$,

$$\mathcal{M}_n = \mathbb{E}\left[\varphi(X_T) \mid X_n\right] \ .$$

Moreover, since X_n is a Markov chain, the above conditional expectation is equivalent to the one with respect to $\mathcal{F}_n$. Indeed, $\mathbb{E}\left[\varphi(X_T) \mid \mathcal{F}_{T-1}\right]$ is necessarily of the form $\psi_T(X_0, \dots X_{T-1})$, where ψ_T is measurable and given by: $\psi_T(x_0, \dots, x_{T-1}) = \mathbb{E}\left[\varphi(X_T) \mid X_0 = x_0, \dots, X_{T-1} = x_{T-1}\right]$ and since X_n is a Markov chain, we get that $\psi_T(x_0, \dots, x_{T-1}) = \mathbb{E}\left[\varphi(X_T) \mid X_{T-1} = x_{T-1}\right]$ depends only on x_{T-1}, so $\mathbb{E}\left[\varphi(X_T) \mid \mathcal{F}_{T-1}\right] = \psi_T(X_{T-1})$. By induction, we get also that $\mathbb{E}\left[\varphi(X_T) \mid \mathcal{F}_n\right]$ is a measurable function of X_n. Since the σ-algebra generated by X_n is smaller than $\mathcal{F}_n$ (which is generated by $X_0, \dots, X_n$), we deduce that $\mathbb{E}\left[\varphi(X_T) \mid \mathcal{F}_n\right] = \mathbb{E}\left[\varphi(X_T) \mid X_n\right]$.

Therefore,

$$\mathcal{M}_n = \mathbb{E}\left[\varphi(X_T) \mid \mathcal{F}_n\right]$$

which is clearly a Martingale, by the property of compositions of conditional expectations:

$$\mathbb{E}\left[\mathcal{M}_{n+1} \mid \mathcal{F}_n\right] = \mathbb{E}\left[\mathbb{E}\left[\varphi(X_T) \mid \mathcal{F}_{n+1}\right] \mid \mathcal{F}_n\right] = \mathbb{E}\left[\varphi(X_T) \mid \mathcal{F}_n\right] = \mathcal{M}_n \ .$$

This implies that

$$v_k(X_k) = \mathcal{M}_k = \mathbb{E}\left[\mathcal{M}_{k+1} \mid \mathcal{F}_k\right] = \mathbb{E}\left[v_{k+1}(X_{k+1}) \mid X_k\right] \ ,$$

which gives the recurrence equation of Proposition 2.16.

The following more general result will be used in what follows in order to establish the Bellman dynamic programming equation for Markov decision problems, which will be a nonlinear extension of Kolmogorov equation.

Theorem 2.18 (Kolmogorov Equation for an additive functional)**.** *Let $(X_n)_{n\in\mathbb{N}}$ be a Markov chain over $(\Omega, \mathfrak{A}, P)$, taking its values in $\mathcal{E}$, with Markov transition matrices $M^{(n)}$ at time $n \in \mathbb{N}$. Let $\varphi \in \mathbb{R}^{\mathcal{E}}$, $T \in \mathbb{N}$, and $r_k \in \mathbb{R}^{\mathcal{E}}$ for $0 \leq k \leq T-1$, and denote:*

$$v_k(x) = \mathbb{E}\left[\left(\sum_{\ell=k}^{T-1} r_\ell(X_\ell)\right) + \varphi(X_T) \mid X_k = x\right] \ .$$

Then, v_k satisfies the following backward recurrence equation, called Kolmogorov equation *:*

$$v_k = r_k + M^{(k)} v_{k+1},\ 0 \leq k \leq T-1 \ , \tag{2.3a}$$

with final condition:

$$v_T = \varphi \ . \tag{2.3b}$$

Proof. Use the second way of proof of previous Kolmogorov equation:

$$\begin{aligned}
v_k(x) &= \mathbb{E}\left[\left(\sum_{\ell=k}^{T-1} r_\ell(X_\ell)\right) + \varphi(X_T) \mid X_k = x\right] \\
&= \mathbb{E}\left[r_k(X_k) \mid X_k = x\right] + \mathbb{E}\left[\left(\sum_{\ell=k+1}^{T-1} r_\ell(X_\ell)\right) + \varphi(X_T) \mid X_k = x\right] \\
&= r_k(x) + \mathbb{E}\left[\mathbb{E}\left[\left(\sum_{\ell=k+1}^{T-1} r_\ell(X_\ell)\right) + \varphi(X_T) \mid X_{k+1}\right] \mid X_k = x\right] \\
&= r_k(x) + \mathbb{E}\left[v_{k+1}(X_{k+1}) \mid X_k = x\right] \\
&= r_k(x) + \sum_{y\in\mathcal{E}} P(X_{k+1} = y \mid X_k = x) v_{k+1}(y) \\
&= r_k(x) + \sum_{y\in\mathcal{E}} M^{(k)}_{xy} v_{k+1}(y) \\
&= (r_k + M^{(k)} v_{k+1})_x \ ,
\end{aligned}$$

hence $v_k = r_k + M^{(k)} v_{k+1}$. □

Remark 2.19. As in Remark 2.17, another way to prove Theorem 2.18 is to show that for all $k \geq 0$, the sequence

$$\mathcal{M}_n = \left(\sum_{\ell=k}^{n-1} r_\ell(X_\ell) \right) + v_n(X_n), \quad n \geq k,$$

is a Martingale for the filtration $(\mathcal{F}_n)_{n\geq 0}$ associated to the Markov chain $(X_n)_{n\geq k}$. Indeed, using the definition of v_n, we get, for $n \geq k$,

$$\mathcal{M}_n = \left(\sum_{\ell=k}^{n-1} r_\ell(X_\ell) \right) + \mathbb{E}\left[\left(\sum_{\ell=n}^{T-1} r_\ell(X_\ell) \right) + \varphi(X_T) \mid X_n \right]$$

and since $X_k, \ldots, X_{n-1}$ are measurable with respect to $\mathcal{F}_n$ and X_n is a Markov chain, so the above conditional expectation with respect to X_n is equivalent to the one with respect to $\mathcal{F}_n$, we get

$$\mathcal{M}_n = \mathbb{E}\left[\left(\sum_{\ell=k}^{n-1} r_\ell(X_\ell) \right) + \left(\sum_{\ell=n}^{T-1} r_\ell(X_\ell) \right) + \varphi(X_T) \mid \mathcal{F}_n \right] = \mathbb{E}\left[\left(\sum_{\ell=k}^{T-1} r_\ell(X_\ell) \right) + \varphi(X_T) \mid \mathcal{F}_n \right]$$

which is clearly a Martingale. This implies that

$$v_k(X_k) = \mathcal{M}_k = \mathbb{E}\left[\mathcal{M}_{k+1} \mid \mathcal{F}_k\right] = r_k(X_k) + \mathbb{E}\left[v_{k+1}(X_{k+1}) \mid \mathcal{F}_k\right] = r_k(X_k) + \mathbb{E}\left[v_{k+1}(X_{k+1}) \mid X_k\right] \ ,$$

which gives the recurrence equation of Theorem 2.18.

Remark 2.20. When $r_k = r$ and $M^{(k)} = M$ do not depend on k, (hence the Markov chain $(X_n)_{n\in\mathbb{N}}$ is stationary), the Kolmogorov equation writes $v_k = r + Mv_{k+1}$, so that one can consider the value function as a function of the remaining time until the end:

$$v^{(t)}(x) = \mathbb{E}\left[\left(\sum_{\ell=0}^{t-1} r_\ell(X_\ell) \right) + \varphi(X_t) \mid X_0 = x \right] \ ,$$

which satisfies a *forward Kolmogorov* equation:

$$v^{(t+1)} = r + Mv^{(t)} \ .$$

Remark 2.21. Moreover, Kolmogorov equation can be rewritten as

$$v_k - v_{k-1} + (M - I)v_k + r = 0 \ .$$

which is analogue to the Kolmogorov equation of a Markov process (with continuous time):

$$\frac{dv}{dt} + (M - I)v + r = 0 \ .$$

Then, $M - I$ is called the *infinitesimal generator* of the Markov chain. (A Markov process is obtained when the holding times in each state are random independent with exponential law).

A similar Kolmogorov equation is obtained when $(X_t)_{t\geq 0}$ is a Wiener process:

$$\frac{dv}{dt} + \frac{1}{2}\Delta v + r = 0 \ .$$

Exercise 2.5.1. Compute

$$v := \mathbb{E}\left[\left(\sum_{\ell=0}^{T-1} r_\ell(X_\ell)\right) + \varphi(X_T)\right] \ ,$$

for any initial law of the Markov chain $(X_n)_{n\geq 0}$.

Exercise 2.5.2. Let X_n be a Markov chain with values in (a finite subset of) $\mathbb{N}$, and consider the sequence $Y_n = X_0 + \cdots + X_n$. Compute

$$v(x) = \mathbb{E}\left[\varphi(Y_T) \mid X_0 = x\right] \ .$$

Do the same for $Y_n = \max(X_0, \ldots, X_n)$.

Exercise 2.5.3. Consider a Markov chain $(X_n)_{n\geq 0}$ with values in $\mathcal{E}$ and transition matrix $M \in \mathbb{R}^{\mathcal{E}\times\mathcal{E}}$ (independent of time). Let f be a function from $\mathcal{E}$ to $\mathbb{R}$, compute

$$v_T(x) = \mathbb{P}\left(\exists n \in \{0, \ldots, T\},\ f(X_n) \geq 1 \mid X_0 = x\right) \ .$$

Taking the exponential of an additive functional does not change optimization problems, but it does change expectation.

Theorem 2.22 (Kolmogorov equation for a multiplicative functional)**.** *Let $(X_n)_{n\in\mathbb{N}}$ be a Markov chain over $(\Omega, \mathfrak{A}, P)$, taking its values in $\mathcal{E}$, with Markov transition matrices $M^{(n)}$ at time $n \in \mathbb{N}$. Let $\varphi \in \mathbb{R}^{\mathcal{E}}$, $T \in \mathbb{N}$, and $\alpha_k \in \mathbb{R}_+^{\mathcal{E}}$, for $0 \leq k \leq T-1$, and denote:*

$$v_k(x) = \mathbb{E}\left[\left(\prod_{\ell=k}^{T-1} \alpha_\ell(X_\ell)\right)\varphi(X_T) \mid X_k = x\right] \ .$$

Let $A^{(k)} \in \mathbb{R}^{\mathcal{E}\times\mathcal{E}}$ be the matrix with nonnegative entries $A^{(k)}_{xy} = \alpha_k(x)M^{(k)}_{xy}$, for $x, y \in \mathcal{E}$. Then, v_k satisfies the following backward recurrence equation:

$$v_k = A^{(k)} v_{k+1},\ 0 \leq k \leq T-1 \ , \tag{2.4a}$$

with final condition:

$$v_T = \varphi \ . \tag{2.4b}$$

Proof. We use again the same arguments as for the previous Kolmogorov equations:

$$\begin{aligned}
v_k(x) &= \mathbb{E}\left[\left(\prod_{\ell=k}^{T-1} \alpha_\ell(X_\ell)\right)\varphi(X_T) \mid X_k = x\right] \\
&= \alpha_k(x)\mathbb{E}\left[\mathbb{E}\left[\left(\prod_{\ell=k+1}^{T-1} \alpha_\ell(X_\ell)\right)\varphi(X_T) \mid X_{k+1}\right] \mid X_k = x\right] \\
&= \alpha_k(x)\mathbb{E}\left[v_{k+1}(X_{k+1}) \mid X_k = x\right] \\
&= \alpha_k(x)\left(\sum_{y\in\mathcal{E}} P(X_{k+1} = y \mid X_k = x) v_{k+1}(y)\right) \\
&= \alpha_k(x)\left(\sum_{y\in\mathcal{E}} M^{(k)}_{xy} v_{k+1}(y)\right) \\
&= \sum_{y\in\mathcal{E}} A^{(k)}_{xy} v_{k+1}(y) \ ,
\end{aligned}$$

which gives $v_k = A^{(k)} v_{k+1}$. □

Again, with the same arguments, one prove:

Theorem 2.23 (Kolmogorov equation for a mixed functional)**.** *Let $(X_n)_{n\in\mathbb{N}}$ be a Markov chain over $(\Omega, \mathfrak{A}, P)$, taking its values in $\mathcal{E}$, with Markov transition matrices $M^{(n)}$ at time $n \in \mathbb{N}$. Let $\varphi \in \mathbb{R}^{\mathcal{E}}$, $T \in \mathbb{N}$, $r_k \in \mathbb{R}^{\mathcal{E}}$, and $\alpha_k \in \mathbb{R}_+^{\mathcal{E}}$, for $0 \le k \le T-1$, and denote:*

$$v_k(x) = \mathbb{E}\left[\left(\sum_{\ell=k}^{T-1}\left(\prod_{m=k}^{\ell-1}\alpha_m(X_m)\right) r_\ell(X_\ell)\right) + \left(\prod_{m=k}^{T-1}\alpha_m(X_m)\right)\varphi(X_T) \mid X_k = x\right] \ .$$

Let $A^{(k)} \in \mathbb{R}^{\mathcal{E}\times\mathcal{E}}$ be the matrix with nonnegative entries $A^{(k)}_{xy} = \alpha_k(x) M^{(k)}_{xy}$, for $x, y \in \mathcal{E}$. Then, v_k satisfies the following backward recurrence equation:

$$v_k = r_k + A^{(k)} v_{k+1},\ 0 \le k \le T-1 \ , \tag{2.5a}$$

with final condition:

$$v_T = \varphi \ . \tag{2.5b}$$

When $\alpha_k(x) \equiv \alpha < 1$, α is the *discount factor*, as for determinitic control problems with infinite horizon.

More generally, when $\alpha_k(x) \le 1$ depends on x, we call it a *variable discount factor*.

In that case, the matrix $A^{(k)}$ satisfies $A^{(k)}\mathbf{1} \le \mathbf{1}$. A matrix A with nonnegative entries and such that $A\mathbf{1} \le \mathbf{1}$ is called a *submarkovian matrix*.

When the matrices $A^{(k)}$ are submarkovian, one can reduce the previous problem/functional to a problem with additive criteria, by adding to the state space $\mathcal{E}$ a cemetery point. Indeed, let c denote this cemetery point, assume that $c \notin \mathcal{E}$, and consider $\mathcal{E}' = \mathcal{E} \cup \{c\}$. Let $M'^{(k)} \in \mathbb{R}^{\mathcal{E}'\times\mathcal{E}'}$ be the matrix such that

$$\begin{aligned}
M'^{(k)}_{xy} &= A^{(k)}_{xy}, \quad \text{when } x, y \in \mathcal{E} \\
M'^{(k)}_{cc} &= 1, \\
M'^{(k)}_{cx} &= 0, \quad \forall x \in \mathcal{E} \\
M'^{(k)}_{xc} &= 1 - \alpha_k(x), \quad \forall x \in \mathcal{E} \ ,
\end{aligned}$$

and extend r_k and φ to $\mathcal{E}'$ in r'_k and φ' respectively by mapping c to 0.

Proposition 2.24. *The value function v_k of Theorem 2.23 is the restriction to $\mathcal{E}$ of the value function v'_k obtained in Theorem 2.18 for the matrices $M'^{(k)}$ and functions r'_k et φ'.*

Note that v'_k satisfies necessarily $v'_k(c) = v'_{k+1}(c) = \cdots = \varphi'(c) = 0$, so the *boundary equation*:

$$v'_k(c) = 0 \ .$$

2.6 Kolmogorov Equations for infinite horizon criteria

Theorem 2.25 (Kolmogorov Equations for a discounted infinite horizon functional)**.** *Let $(X_n)_{n\in\mathbb{N}}$ be a stationary Markov chain over $(\Omega, \mathfrak{A}, P)$, taking its values in $\mathcal{E}$, with Markov transition matrix M. Let $r, \alpha \in \mathbb{R}^{\mathcal{E}}$, satisfying $0 \leq \alpha(x) \leq \bar{\alpha}$ for all $x \in \mathcal{E}$, for some constant $\bar{\alpha} < 1$. Assume that r is bounded in sup-norm (or $\mathcal{E}$ is finite). Denote:*

$$.v(x) = \mathbb{E}\left[\left(\sum_{\ell=0}^{\infty}\left(\prod_{m=0}^{\ell-1}\alpha(X_m)\right)r(X_\ell)\right) \mid X_0 = x\right] \quad . \tag{2.6}$$

Let $A \in \mathbb{R}^{\mathcal{E}\times\mathcal{E}}$ be the matrix with nonnegative entries $A_{xy} = \alpha(x)M_{xy}$, for $x, y \in \mathcal{E}$. Then, v is the unique solution of the fixed point linear equation:

$$v = r + Av \ , \tag{2.7}$$

Before giving the proof let us state some properties of Markov matrices.

Recall that for any matrix $M \in \mathbb{R}^{\mathcal{E}\times\mathcal{E}}$, the *matrix norm* $\|M\|_\infty$ associated to the sup-norm of vectors, defined by:

$$\|M\|_\infty := \sup_{u\in\mathbb{R}^{\mathcal{E}}\setminus\{0\}} \frac{\|Mu\|_\infty}{\|u\|_\infty}, \quad \|u\|_\infty =: \sup_{x\in\mathcal{E}} |u_x| \ ,$$

satisfies

$$\|M\|_\infty = \sup_{x\in\mathcal{E}}\left(\sum_{y\in\mathcal{E}}|M_{xy}|\right) \quad ,$$

and that *spectral radius* of M, denoted $\rho(M)$, which is the maximum of the modulus of its eigenvalues satisfies necessarily $\rho(M) \leq \|M\|$ for any matrix norm.

Lemma 2.26. *For any Markov matrix $M \in \mathbb{R}^{\mathcal{E}\times\mathcal{E}}$, we have*

$$\rho(M) = \|M\|_\infty = 1 \ .$$

Proof. Since a Markov matrix M has nonnegative entries, and the sum of each row is equal to 1, we get $\|M\|_\infty = 1$.

Since 1 is an eigenvalue of M associated to the eigenvector $\mathbf{1}$ ($M\mathbf{1} = \mathbf{1}$), we get $\rho(M) \geq 1$.

Since $1 \leq \rho(M) \leq \|M\|_\infty = 1$, we deduce the result. □

Proof of Theorem 2.25. Assume that $\mathcal{E}$ is finite. Since $\alpha(X_m) \leq \bar{\alpha} < 1$, for all $m \geq 0$, and r is bounded, the series inside the expectation in (2.6) is normally converging (for sup-norm), and its sum is a.s. bounded by $C/(1-\bar{\alpha})$, where C is a bound of r. Hence, the expectation $v(x)$ exists, for all $x \in \mathcal{E}$, which allows one to define the value function $v : x \mapsto v(x)$.

By the same arguments as in previous proofs, one shows that v satisfies (2.7). So it remains to prove that (2.7) has a unique solution. This holds if $I - A$ is invertible, that is 1 is not an eigenvalue of A. Using the formula of the sup-norm of a matrix, we get that $\|A\|_\infty \leq \bar{\alpha}\|M\|_\infty = \bar{\alpha} < 1$, which implies that $\rho(A) \leq \bar{\alpha} < 1$, and so $I - A$ is invertible.

One can also show that the Kolmogorov operator: $\mathcal{K} : v \mapsto r + Av$ is monotone and contracting for the sup-norm with factor $\bar{\alpha} < 1$. Indeed

$$\|\mathcal{K}(v) - \mathcal{K}(w)\|_\infty = \|A(v-w)\|_\infty \leq \|A\|_\infty\|v-w\|_\infty \leq \bar{\alpha}\|v-w\|_\infty \ .$$

So $\mathcal{K}$ has a unique fixed point.

This property holds also when $\mathcal{E}$ is a countable set (discrete infinite), by considering $\mathcal{K}$ as an operator on the Banach space $L^\infty(\mathcal{E})$ of bounded functions from $\mathcal{E}$ to $\mathbb{R}$, endowed with the sup-norm. □

In sustainable development problems, on may wish to consider a functional which put more weight on future rewards. This would mean that some of the discount factors $\alpha(x)$ are greater than 1, and so A is no more contracting for the sup-norm. However, the previous result remains if the spectral radius of A remains lower than 1 as in the following result.

Theorem 2.27. *Let $(X_n)_{n\in\mathbb{N}}$ be a stationary Markov chain over $(\Omega, \mathfrak{A}, P)$, taking its values in a finite set $\mathcal{E}$, with Markov transition matrix M. Let $r, \alpha \in \mathbb{R}^{\mathcal{E}}$, satisfying $0 \leq \alpha(x)$ for all $x \in \mathcal{E}$. Let $A \in \mathbb{R}^{\mathcal{E}\times\mathcal{E}}$ be the matrix with nonnegative entries $A_{xy} = \alpha(x)M_{xy}$, for $x, y \in \mathcal{E}$. Assume that the spectral radius $\bar{\alpha} := \rho(A)$ is < 1. Then, for all $x \in S$, the value $v(x)$ of (2.6) is well defined and the function $v : \mathcal{E} \to \mathbb{R}$, $x \in \mathcal{E} \mapsto v(x)$ is the unique solution of the fixed point linear equation (2.7).*

Proof. Since S is a finite set, r is bounded by some constant C. Hence

$$\sum_{\ell=0}^{\infty} \left| \left(\prod_{m=0}^{\ell-1} \alpha(X_m) \right) r(X_\ell) \right| \leq CY$$

where Y is the nonnegative random variable defined by

$$Y := \sum_{\ell=0}^{\infty} \left(\prod_{m=0}^{\ell-1} \alpha(X_m) \right) \ .$$

Since Y is the sum of series with nonnegative coeffcients, its expectation exists and is equal to

$$\mathbb{E}\left[Y \mid X_0 = x\right] = \sum_{\ell=0}^{\infty} w_k(x) \quad \text{with} \quad w_k(x) := \mathbb{E}\left[\prod_{m=0}^{k-1} \alpha(X_m) \mid X_0 = x \right] \ .$$

Using Kolmogorov equation for multiplicative functionals, we obtain that w_k satisfies $w_k = Aw_{k-1}$ and $w_0 = \mathbf{1}$, hence $w_k = A^k\mathbf{1}$. Therefore, $\mathbb{E}\left[Y \mid X_0 = x\right] = \sum_{\ell=0}^{\infty}(A^k\mathbf{1})_x$. Since $\rho(A) < 1$, the previous series converges and is equal to $((I - A)^{-1}\mathbf{1})_x$. Then, Y has a finite expectation, which implies (by dominated convergence theorem) that the random variable $\sum_{\ell=0}^{\infty}\left(\prod_{m=0}^{\ell-1}\alpha(X_m)\right) r(X_\ell)$ is well defined (exists almost surely) and has a finite expectation. Hence, the value $v(x)$ in (2.6) is well defined. The rest of the result can be proved using the same arguments as for previous theorem. □

2.7 Kolmogorov Equations for stopping time criteria

Definition 2.28. Given a filtration $(\mathcal{F}_n)_{n\in\mathbb{N}}$ on $(\Omega, \mathfrak{A})$, a random variable τ with values in $\mathbb{N}\cup\{+\infty\}$ is a *stopping time* with respect to $(\mathcal{F}_n)_{n\in\mathbb{N}}$ if $\{\tau = n\} \in \mathcal{F}_n$, for all $n \in \mathbb{N}$. We then denote

$$\mathcal{F}_\tau := \{A \in \mathfrak{A} \mid \{\tau \leq t\} \cap A \in \mathcal{F}_t,\ \forall t \geq 0\} \ .$$

Fact 2.29. τ is a stopping time if and only if $\{\tau \leq n\} \in \mathcal{F}_n$, for all $n \in \mathbb{N}$.

Proof. If τ is a stopping time, then $\{\tau \leq n\} = \cup_{k=0,\dots,n}\{\tau = k\} \in \mathcal{F}_n$, since $\mathcal{F}_k \subset \mathcal{F}_n$, for all $k \leq n$. Conversely, if $\{\tau \leq n\} \in \mathcal{F}_n$, for all $n \in \mathbb{N}$, then $\{\tau = n\} = \{\tau \leq n\} \setminus \{\tau \leq n-1\} \in \mathcal{F}_n$, since $\mathcal{F}_{n-1} \subset \mathcal{F}_n$. □

Definition 2.30. Given a Markov chain $(X_n)_{n\in\mathbb{N}}$ over $(\Omega, \mathfrak{A}, P)$, we associate the filtration $(\mathcal{F}_n)_{n\in\mathbb{N}}$:

$$\mathcal{F}_n := \sigma^{\mathrm{a}}(X_0, \dots, X_n), \quad \forall n \in \mathbb{N} \ .$$

Then, a random variable τ with values in $\mathbb{N} \cup \{+\infty\}$ is a *stopping time* with respect to the Markov chain $(X_n)_{n\in\mathbb{N}}$ if it is a stopping time with respect to $(\mathcal{F}_n)_{n\in\mathbb{N}}$.

Note that the filtration associated to a Markov chain $(X_n)_{n\geq 0}$ is the minimal filtration such that $(X_n)_{n\geq 0}$ is adapted to it.

Example 2.31. Let $(X_n)_{n\in\mathbb{N}}$ be a Markov chain over $(\Omega, \mathfrak{A}, P)$, and let B be a subset of $\mathcal{E}$. Then, for all $k \in \mathbb{N}$,

$$\tau_B^k := \inf\{n \geq k \mid X_n \notin B\}$$

is a stopping time with respect to the Markov chain $(X_n)_{n\geq k}$. Indeed, for $t \geq 0$, $\{\tau_B^k > t\} = \bigcap_{k\leq \ell \leq t}\{X_\ell \in B\}$ belongs to $\sigma^{\mathrm{a}}(X_k, \dots, X_t)$, so does its complementary $\{\tau_B^k \leq t\}$. The stopping time τ_B^k is called the *exit time* from B of the Markov chain starting at time k.

Theorem 2.32 (Strong Markov property)**.** *Let $(X_n)_{n\in\mathbb{N}}$ be a Markov chain over $(\Omega, \mathfrak{A}, P)$, taking its values in $\mathcal{E}$. Let τ be a stopping time with respect to $(X_n)_{n\geq 0}$. We have,*

$$\begin{aligned} &P(X_{\tau+1} = y_1, \dots, X_{\tau+k} = y_k \mid X_0 = x_0, \dots, X_{\tau-1} = x_{\tau-1}, X_\tau = y_0 \text{ and } \tau < +\infty) \\ &\quad = P(X_{\tau+1} = y_1, \dots, X_{\tau+k} = y_k \mid X_\tau = y_0 \text{ and } \tau < +\infty) \ . \end{aligned}$$

for all $k, \ell \in \mathbb{N}$, sequences $(x_\ell)_{\ell\geq 0}$, and $y_0, \dots, y_k \in \mathcal{E}$.

Moreover if the chain is stationary then

$$P(X_{\tau+1} = y_1, \dots, X_{\tau+k} = y_k \mid X_\tau = y_0 \text{ and } \tau < +\infty) = P(X_1 = y_1, \dots, X_k = y_k \mid X_0 = y_0) \ .$$

Theorem 2.33 (Kolmogorov Equation for a finite horizon functional with stopping time)**.** *Let $(X_n)_{n\in\mathbb{N}}$ be a Markov chain over $(\Omega, \mathfrak{A}, P)$, taking its values in $\mathcal{E}$, with Markov transition matrices $M^{(n)}$ at time $n \in \mathbb{N}$. Let $\varphi \in \mathbb{R}^{\mathcal{E}}$, $B \subset \mathcal{E}$ be nonempty, $T \in \mathbb{N}$, and $r_k \in \mathbb{R}^B$ for $0 \leq k \leq T-1$, and denote, for all $x \in \mathcal{E}$:*

$$v_k(x) = \mathbb{E}\left[\left(\sum_{\ell=k}^{T\wedge\tau_B^k - 1} r_\ell(X_\ell)\right) + \varphi(X_{T\wedge\tau_B^k}) \mid X_k = x\right] \ .$$

Then, v_k satisfies the following backward recurrence equation, called Kolmogorov equation *:*

$$v_k(x) = r_k(x) + (M^{(k)} v_{k+1})(x), \ x \in B, \ 0 \leq k \leq T-1 \ , \tag{2.8a}$$

with boundary condition

$$v_k(x) = \varphi(x), \ x \notin B \ , \tag{2.8b}$$

and final condition

$$v_T(x) = \varphi(x), \ x \in B \ . \tag{2.8c}$$

Note that here, we used the same name φ for the function involved in the boundary condition and in the final condition. In general, and in particular when the state space is continuous, one consider two maps $\varphi \in \mathbb{R}^B$ and $\psi \in \mathbb{R}^{\mathcal{E}\setminus B}$, and define v_k as the functional

$$v_k(x) = \mathbb{E}\left[\left(\sum_{\ell=k}^{T\wedge\tau_B^k-1} r_\ell(X_\ell)\right) + \varphi(X_T)\mathbf{1}_{T<\tau_B^k} + \psi(X_{\tau_B^k})\mathbf{1}_{T\geq\tau_B^k} \mid X_k = x\right] \ .$$

Proof of Theorem 2.33. Consider the sequence $Y_n = X_{n\wedge\tau_B^k}$, for $n \geq k$. With the same arguments as for the strong Markov property, we can show that this is a Markov chain starting at time $n = k$. Indeed,

$$\begin{aligned}
P(Y_n = y_n, \cdots, Y_k = y_k) =& \sum_{p=k}^{n-1} P(Y_n = y_n, \cdots, Y_k = y_k, \tau_B^k = p) \\
&+ P(Y_n = y_n, \cdots, Y_k = y_k, \tau_B^k \geq n) \\
=& \sum_{p=k}^{n-1} P(X_{n\wedge p} = y_n, \cdots, X_k = y_k, \tau_B^k = p) \\
&+ P(X_n = y_n, \cdots, X_k = y_k, \tau_B^k \geq n) \\
=& \sum_{p=k}^{n-1} P(X_p = y_p, \cdots, X_k = y_k)\mathbf{1}_{y_k\in B}\cdots\mathbf{1}_{y_{p-1}\in B}\mathbf{1}_{y_p\notin B}\mathbf{1}_{y_p=y_{p+1}=\cdots=y_n} \\
&+ P(X_n = y_n, \cdots, X_k = y_k)\mathbf{1}_{y_k\in B}\cdots\mathbf{1}_{y_{n-1}\in B} \\
=& \sum_{p=k}^{n=1} M^{(p-1)}_{y_{p-1}y_p}\cdots M^{(k)}_{y_k y_{k+1}}\mathbf{1}_{y_k\in B}\cdots\mathbf{1}_{y_{p-1}\in B}\mathbf{1}_{y_p\notin B}\mathbf{1}_{y_p=y_{p+1}=\cdots=y_n} \\
&+ M^{(n-1)}_{y_{n-1}y_n}\cdots M^{(k)}_{y_k y_{k+1}}\mathbf{1}_{y_k\in B}\cdots\mathbf{1}_{y_{n-1}\in B} \\
=& M^{(B,n-1)}_{y_{n-1}y_n}\cdots M^{(B,k)}_{y_k y_{k+1}}
\end{aligned}$$

where, for $n \geq k$, $M^{(B,n)}$ is given by:

$$M^{(B,n)}_{xy} = \begin{cases} M^{(n)}_{xy} & \text{for } x \in B, y \in \mathcal{E} \\ 1 & \text{for } x = y \notin B \\ 0 & \text{otherwise.} \end{cases}$$

Hence, by Proposition 2.3, $(Y_n)_{n\geq k}$ is a Markov chain with transition matrix $M^{(B,n)}$ at time $n \geq k$.

For all $n \geq 0$, extend r_n by 0 on $\mathcal{E} \setminus B$. Then, for all $x \in \mathcal{E}$, v_k coincides with:

$$v_k(x) = \mathbb{E}\left[\left(\sum_{\ell=k}^{T-1} r_\ell(Y_\ell)\right) + \varphi(Y_T) \mid Y_k = x\right] \ .$$

From Theorem 2.18, we get the recurrence equation

$$v_k = r_k + M^{(B,k)}v_{k+1},\ 0 \leq k \leq T-1 \ ,$$

with final condition: $v_T = \varphi$. This final condition implies (2.8c). When $x \in B$, the recurrence equation gives (2.8a). When $x \notin B$, the recurrence equation gives $v_k(x) = v_{k+1}(x)$, which with the final condition implies $v_k(x) = \varphi(x)$, hence the boundary equation (2.8b). □

Another way to prove Theorem 2.33 is to use Theorem 2.23 and the following result, which is easy to check.

Fact 2.34. The value function v_k of Theorem 2.33 can rewritten as the value function of the mixed functional :

$$v_k(x) = \mathbb{E}\left[\left(\sum_{\ell=k}^{T-1}\left(\prod_{m=k}^{\ell-1}\alpha_m(X_m)\right)r'_\ell(X_\ell)\right) + \left(\prod_{m=k}^{T-1}\alpha_m(X_m)\right)\varphi(X_T) \mid X_k = x\right] \ ,$$

for the same Markov chain $(X_n)_{n\geq 0}$, with the same final reward φ, and the instantaneous rewards r'_k and variable discount factors α_k given by:

$$\begin{aligned} r'_k(x) &= r_k(x), \quad \text{for } x \in B \\ r'_k(x) &= \varphi(x), \quad \text{for } x \notin B \\ \alpha_k(x) &= 1, \quad \text{for } x \in B \\ \alpha_k(x) &= 0, \quad \text{for } x \notin B \ . \end{aligned}$$

We can derive similarly the solution of the discounted infinite horizon problem with stopping time.

Theorem 2.35 (Kolmogorov Equations for discounted infinite horizon with stopping time)**.** *Let $(X_n)_{n\in\mathbb{N}}$ be a stationary Markov chain over $(\Omega, \mathfrak{A}, P)$, taking its values in $\mathcal{E}$, with Markov transition matrix M. Let $B \subset \mathcal{E}$ be nonempty, $r \in \mathbb{R}^B$, and $\alpha \in (0,1)$, Assume that r is bounded in sup-norm (or $\mathcal{E}$ is finite). Denote, for all $x \in \mathcal{E}$,*

$$v(x) = \mathbb{E}\left[\left(\sum_{\ell=0}^{\tau_B-1}\alpha^\ell r(X_\ell)\right) + \alpha^{\tau_B}\varphi(X_{\tau_B}) \mid X_0 = x\right] \ . \tag{2.9}$$

Then, v is the unique solution of the fixed point linear equation:

$$\left\{\begin{array}{lll} v(x) & = r(x) + \alpha(Mv)(x) & x \in B \ , \\ v(x) & = \varphi(x) & x \notin B \ . \end{array}\right.$$

We can in some cases avoid the discount factor.

Theorem 2.36 (Kolmogorov Equations for undiscounted infinite horizon with stopping time)**.** *Let $(X_n)_{n\in\mathbb{N}}$ be a stationary Markov chain over $(\Omega, \mathfrak{A}, P)$, taking its values in $\mathcal{E}$, with Markov transition matrix M. Let $B \subset \mathcal{E}$ be nonempty, $r \in \mathbb{R}^B$. Assume that r is bounded in sup-norm (or $\mathcal{E}$ is finite), and that the matrix $M_{BB} \in \mathbb{R}^{B\times B}$, which is the restriction of M to rows and columns in B ($(M_{BB})_{xy} = M_{xy}$ for all $x, y \in B$) has a spectral radius $\rho < 1$. Denote, for all $x \in \mathcal{E}$,*

$$v(x) = \mathbb{E}\left[\left(\sum_{\ell=0}^{\tau_B-1} r(X_\ell)\right) + \varphi(X_{\tau_B}) \mid X_0 = x\right] \ . \tag{2.10}$$

Then, v is well defined and it is the unique solution of the fixed point linear equation:

$$\left\{\begin{array}{lll} v(x) & = r(x) + (Mv)(x) & x \in B \ , \\ v(x) & = \varphi(x) & x \notin B \ . \end{array}\right.$$

Proof. Same arguments as for the proof of Theorem 2.27. □

Example 2.37. Consider a Markov chain with transition matrix

$$M = \begin{bmatrix} 1/2 & 1/2 & 0 \\ 1/2 & 1/2 & 0 \\ 0 & 1/2 & 1/2 \end{bmatrix} .$$

If one consider $B = \{1, 2\} \subset \mathcal{E}$, then $\tau_B = +\infty$ almost surely since B is a recurrence class. Then, $\mathbb{E}\left[(\sum_{k=0}^{\tau_B - 1} r(X_k)) + \varphi(X_{\tau_B}) \mid X_0 = x\right] = \infty$ if for instance $r = (1\ 1\ 0)^T$.

Now if $B = \{2, 3\}$, then $\tau_B < +\infty$ almost surely, and

$$M_{BB} = \begin{bmatrix} 1/2 & 0 \\ 1/2 & 1/2 \end{bmatrix}$$

satisfies $\rho(M_{BB}) = 1/2$. Then, $v(x) = \mathbb{E}\left[(\sum_{k=0}^{\tau_B - 1} r(X_k)) + \varphi(X_{\tau_B}) \mid X_0 = x\right]$ exists and is solution of

$$\begin{aligned} v(1) &= \varphi(1) \\ v(2) &= r(2) + \frac{1}{2}v(1) + \frac{1}{2}v(2) \\ v(3) &= r(3) + \frac{1}{2}v(2) + \frac{1}{2}v(3) \ . \end{aligned}$$

This gives

$$\begin{aligned} v(1) &= \varphi(1) \\ v(2) &= 2r(2) + \varphi(1) \\ v(3) &= 2r(3) + 2r(2) + \varphi(1) \ . \end{aligned}$$

2.8 Further examples

Example 2.38 (A rolling dice game)**.** Consider a game in N steps. At each step, the player is rolling a dice. If the dice falls on 6, the player is loosing all his previous gains positive or negative, otherwise he receives the value of the dice minus 3 as an additional gain. Denoting by W_n the value of the dice at the nth stage of the game, we get that $(W_n)_{n\geq 0}$ is an i.i.d. sequence of random variables with laws: $P(W_n = i) = 1/6$ for $i \in \{1, \ldots, 6\}$.

If the dice neither fall on 6, the total gain/payoff of the player at stage n would be equal to $W_1 - 3 + \cdots + W_n - 3$. Since the player may loose everything, one need to consider a new sequence X_n of states consisting in the (possible) total reward at time n. We starts with $X_0 = 0$ and get that

$$X_{n+1} = \begin{cases} X_n + W_{n+1} - 3 & \text{if } W_{n+1} \neq 6 \\ 0 & \text{otherwise,} \end{cases}$$

for all $n = 0, \ldots, N-1$. Such a dynamics can be written $X_{n+1} = f(X_n, W_{n+1})$, which implies that the sequence $(X_n)_{n\geq 0}$ is a Markov chain with values in $\mathbb{Z}$. The expected total reward at the end of the game is then equal to $\mathbb{E}[X_T]$, which has the form of the criterion considered in Proposition 2.16.

Then it can be computed by using the Kolmogorov equation. Using the dynamics f, instead of transition probabilities, Kolmogorov equation can be written as follows

$$v_k(x) = \mathbb{E}\left[v_{k+1}(f(x, W_{k+1}))\right], \quad v_T(x) = x, \quad x \in \mathbb{Z} \ .$$

Then, $\mathbb{E}[X_T] = v_0(0)$ (since $X_0 = 0$). Using the above informations on f and the law of $(W_n)_{n\geq 0}$, we can rewrite the Kolmogorov equation as:

$$v_k(x) = \frac{1}{6}\{v_{k+1}(0) + \sum_{i=1}^{5} v_{k+1}(x+i-3)\} \quad x \in \mathbb{Z} \ .$$

Then, using that $v_T(x) = x$, we can prove by backward induction on k that v_k is linear in x: $v_k(x) = z_k x$ with $z_k = z_{k+1}\frac{5}{6}$. This implies that the game has no expected gain.

2.9 Solutions of Exercises

Exercise 2.3.1.

Exercise 2.5.1. Let

$$v := \mathbb{E}\left[\left(\sum_{\ell=0}^{T-1} r_\ell(X_\ell)\right) + \varphi(X_T)\right] \ .$$

Using the property that $v = \mathbb{E}\left[\mathbb{E}\left[\left(\sum_{\ell=0}^{T-1} r_\ell(X_\ell)\right) + \varphi(X_T) \mid X_0\right]\right]$, we get that $v = p^{(0)}v_0$, where v_0 is the solution of Kolmogorov equation.

Exercise 2.5.2. Let X_n be a Markov chain with values in (a finite subset $\mathcal{E}$ of) $\mathbb{N}$, and consider the sequence $Y_n = X_0 + \cdots + X_n$. Then, Y_n satisfies the recurrence $Y_{n+1} = Y_n + X_{n+1}$. Since X_n is a Markov chain, then (X_n, Y_n) is a Markov chain on $\mathcal{E} \times \mathbb{N}$. Let $\tilde{M}^{(n)}$ be its transition probability matrix. We have

$$\tilde{M}^{(n)}_{(x,y)(x',y')} = M^{(n)}_{xx'}\mathbf{1}_{(y+x')y'} \ .$$

Moreover

$$v(x) = \mathbb{E}\left[\varphi(Y_T) \mid X_0 = x\right] = \mathbb{E}\left[\tilde{\varphi}(X_T, Y_T) \mid (X_0, Y_0) = (x, 0)\right] \ ,$$

with $\tilde{\varphi}(x, y) = \varphi(y)$. So $v(x) = w_0(x, 0)$ where w_k satisfies Kolmogorov equation $w_k = \tilde{M}^{(k)}w_{k+1}$, that is

$$w_k(x, y) = \sum_{x' \in \mathcal{E}} M^{(k)}_{xx'} w_{k+1}(x', y + x') \ ,$$

with $w_T = \tilde{\varphi}$.

Exercise 2.5.3. Consider a Markov chain $(X_n)_{n\geq 0}$ with values in $\mathcal{E}$ and transition matrix $M \in \mathbb{R}^{\mathcal{E}\times\mathcal{E}}$ (independent of time). Let f be a function from $\mathcal{E}$ to $\mathbb{R}$, compute

$$v_T(x) = \mathbb{P}\left(\exists n \in \{0, \ldots, T\},\ f(X_n) \geq 1 \mid X_0 = x\right) \ .$$

Chapter 3

Markov decision processes with finite horizon criteria

We now consider a discrete time dynamical system $(X_n)_{n\geq 0}$ with finite (or discrete) state space $\mathcal{E}$ and a dynamics of the following type:

$$X_{n+1} = f_n(X_n), \quad n \geq 1 \ ,$$

which can both be changed (by a company manager, an investor, a provider, a driver,...), by applying an *action* or *control* U_n at each time or stage $n \geq 0$, and be subject to randomness. For instance:

$$X_{n+1} = f_n(X_n, U_n, W_n), \quad n \geq 0 \ .$$

The aim is still to choose the sequence of actions $U_0, \ldots, U_k, \ldots$ in such a way that they minimize (resp. maximize) a certain functional, called the total cost (resp. the total payoff).

However, we assume that information on the sequences X_n and W_n arrive sequentially, so that at time n, the "manager" only knows $X_0, \ldots, X_n$ and $W_0, \ldots, W_n$. Then, the decision to choose U_n is taken at time n using this information. We are still in the context of *complete observation* since we know all the past states, however we do not know the future states. This is the main difference with deterministic control problems, in which, given the model, and the state at some time n, we can infer all the state trajectory $(X_k)_{k\geq n}$.

Since we cannot observe the future realizations of the Markov chain (when the sequence of actions is fixed for instance), we shall optimze a functional which is the expectation of the total payoff given the initial state, which is a criteria already considered in the chapter on Markov chains, Chapter 2.

Another difficulty is that the choice of the actions $(U_n)_{n\geq 0}$ changes the random process $(X_n)_{n\geq 0}$, so we cannot start with a Markov chain on a given probability space as in previous chapter. We can only define a model with all the parameters of the system, like the dynamics or the transition probabilities, and the initial law. This is what is called a *Markov decision process* or a *controlled Markov chain*.

3.1 Markov decision processes

Definition 3.1. A *Markov Decision Process (MDP)* or a *controlled Markov chain* consists in giving the following parameters:

- a finite or discrete *state space* $\mathcal{E}$;
- an *action space* $\mathcal{C}$
- for all $k \in \mathbb{N}$ and $x \in \mathcal{E}$, the subset $\mathcal{C}_k(x) \subset \mathcal{C}$ of all possible actions at time k, when the state is equal to x;
- for all $k \in \mathbb{N}$, the set $\mathcal{A}_k := \{(x,u) \mid x \in \mathcal{E},\ u \in \mathcal{C}_k(x)\}$ of all possibles couples (state, action) at time k;
- an initial probability $p^{(0)} \in \Delta_{\mathcal{E}}$ on $\mathcal{E}$, or an initial state $x_0 \in \mathcal{E}$, which is equivalent to the case where $p^{(0)}$ is the Dirac measure at x_0;
- for all $k \in \mathbb{N}$, $x \in \mathcal{E}$ and $u \in \mathcal{C}_k(x)$, a probability row vector $M_x^{(k,u)}$ over $\mathcal{E}$, the entries of which will be denoted $\left(M_{xy}^{(k,u)}\right)_{y\in\mathcal{E}}$.

The MDP is *stationary* if $\mathcal{C}_k(x)$ and $M_x^{(k,u)}$ do not depend on time k. In this case, the index or argument k is omitted. It is *uncontrolled* if the sets $\mathcal{C}_k(x)$ are singletons. In this case, the argument u is omitted.

In the uncontrolled case, the above parameters allow one to construct a Markov chain (and a probability space), by considering the transition probability matrices

$$M_{xy}^{(k)} = P(X_{k+1} = y \mid X_k = x) \ .$$

In the general case, one wish to construct (a probability space and) two discrete time processes $(X_k)_{k\geq 0}$ and $(U_k)_{k\geq 0}$ taking their values in $\mathcal{E}$ and $\mathcal{C}$ respectively, with transition probabilities $M_{xy}^{(k,u)}$:

$$M_{xy}^{(k,u)} = P(X_{k+1} = y \mid X_k = x, U_k = u) \ , \tag{3.1a}$$

and such that (X_k) satisfies the following Markov property:

$$\begin{aligned} &P(X_{k+1} = x_{k+1} \mid X_k = x_k, U_k = u_k, X_{k-1} = x_{k-1}, U_{k-1} = u_{k-1}, \dots, X_0 = x_0, U_0 = u_0) \\ &\quad = \quad P(X_{k+1} = x_{k+1} \mid X_k =_k, U_k = u_k) \ , \quad \forall x_i \in \mathcal{E},\ u_i \in \mathcal{C}_i(x_i), \text{ for } i \geq 0 \ . \end{aligned} \tag{3.1b}$$

To define the underlying probability, one need to assume that the control process $(U_k)_{k\geq 0}$ is admissible, meaning that U_k depends only on the past of states and actions, or more precisely that $(U_k)_{k\geq 0}$ is obtained from a strategy, where we extend the notion of strategy as follows.

Definition 3.2. Given a MDP as above, the set $\mathcal{H}_k = \mathcal{A}_0 \times \cdots \times \mathcal{A}_{k-1} \times \mathcal{E}$ is called the set of *histories* at time k. A *pure strategy* is a sequence $\sigma = (\sigma_k)_{k\geq 0}$ such that, for all $k \geq 0$, σ_k, called the strategy at time k, is a map from $\mathcal{H}_k$ to $\mathcal{C}$ satisfying

$$\sigma_k(x_0, u_0, \dots, x_{k-1}, u_{k-1}, x_k) \in \mathcal{C}_k(x_k), \text{ for all } (x_0, u_0, \dots, x_{k-1}, u_{k-1}, x_k) \in \mathcal{H}_k \ .$$

We denote by Σ the set of all pure strategies. A pure strategy gives rise to the stochastic process $(X_k, U_k)_{k\geq 0}$ with transition probabilities as in (3.1), satisfying in addition

$$U_k = \sigma_k(X_0, U_0, \dots, X_{k-1}, U_{k-1}, X_k) \ ,$$

that is there exists a probability space $(\Omega, \mathfrak{A}, P)$ and a stochastic process $(X_k, U_k)_{k\geq 0}$ over this space satisfying all the above properties. Such a sequence $(X_k, U_k)_{k\geq 0}$ is also called an *admissible sequence* of states and controls.

Definition 3.3. A *random (or relaxed) strategy* is a sequence $\sigma = (\sigma_k)_{k\geq 0}$ such that, for all $k \geq 0$, σ_k is a map from $\mathcal{H}_k$ to the space of probabilities, denoted here $\mathcal{C}^{\mathrm{R}}$, over a given probability space $(\mathcal{C}, \mathfrak{A}_{\mathcal{C}}, P)$ such that the support of $\sigma_k(x_0, u_0, \ldots, x_{k-1}, u_{k-1}, x_k)$ is included in $\mathcal{C}_k(x_k)$, for all $(x_0, u_0, \ldots, x_{k-1}, u_{k-1}, x_k) \in \mathcal{H}_k$. Such a strategy gives rise to a stochastic process $(X_k, U_k)_{k\geq 0}$ satisfying, for all $B \in \mathfrak{A}_{\mathcal{C}}$,

$$P(U_k \in B \mid X_0, U_0, \ldots, X_{k-1}, U_{k-1}, X_k) = [\sigma_k(X_0, U_0, \ldots, X_{k-1}, U_{k-1}, X_k)](B) \ .$$

We denote by Σ^{R} the set of all relaxed strategies.

Definition 3.4. A pure or random strategy is said *Markovian* if each map σ_k depends only on the information on the state at the current time, that is

$$\sigma_k(x_0, u_0, \ldots, x_{k-1}, u_{k-1}, x_k) = \pi_k(x_k) \ ,$$

for some map π_k from S to $\mathcal{C}$ or $\mathcal{C}^{\mathrm{R}}$.

A pure Markovian strategy is also called a *feedback strategy* or *feedback policy*. We denote by $\pi = (\pi_k)_{k\geq 0}$ such a policy and call π_k the policy at time k.

We denote by Π and Π^{R} the sets of all feedback and Markov strategies respectively, and by Π_k and Π_k^{R} the sets of k-coordinates of elements of Π and Π^{R} respectively.

When $\mathcal{C}_k$ and $M_x^{(k,u)}$ do not depend on k (for all $x \in \mathcal{E}$ and $u \in \mathcal{C}(x)$), we say that a (pure or relaxed) Markovian strategy is *stationary* if π_k does not depend on k, in which case π also denotes each of the π_k.

Definition 3.5. A pure strategy is an *open-loop strategy* if σ_k depends only on the initial state, that is

$$\sigma_k(x_0, u_0, \ldots, x_{k-1}, u_{k-1}, x_k) = \omega_k(x_0) \text{ for all } (x_0, u_0, \ldots, x_{k-1}, u_{k-1}, x_k) \in \mathcal{H}_k \ .$$

We denote by OL the set of all open-loop strategies.

Definition 3.6. Given a MDP as in Definition 3.1, and a feedback policy $\pi = (\pi_k)_{k\geq 0} \in \Pi$, we associate the Markov transition matrices $M^{(k,\pi_k)}$ at time k, where for all $k \in \mathbb{N}$, $\pi \in \Pi_k$, the matrix $M^{(k,\pi)}$ is defined by

$$M_{xy}^{(k,\pi)} := M_{xy}^{(k,\pi(x))}, \ \forall x, y \in \mathcal{E} \ .$$

Fact 3.7. Given a MDP as in Definition 3.1, and a feedback policy $\pi = (\pi_k)_{k\geq 0} \in \Pi$, the associated stochastic process $(X_k, U_k)_{k\geq 0}$ as in Definition 3.2 is such that $(X_k)_{k\geq 0}$ is a Markov chain with initial law $p^{(0)}$ and transition probability matrices $M^{(k,\pi_k)}$ at time k. Moreover, $U_k = \pi_k(X_k)$, and so $(X_k, U_k)_{k\geq 0}$ is also a Markov chain taking its values in $\mathcal{E} \times \mathcal{C}$.

Another definition of a MDP is sometimes given, which can be shown to be equivalent to the previous one at least in the case of finite state and action spaces, up to the choice of the probability space.

Definition 3.8. A *Markov Decision Process (MDP)* or a *controlled Markov chain* consists in giving the following parameters:

- a finite or discrete *state space* $\mathcal{E}$;

- an *action space* $\mathcal{C}$
- for all $k \in \mathbb{N}$ and $x \in \mathcal{E}$, the subset $\mathcal{C}_k(x) \subset \mathcal{C}$ of all possible actions at time k, when the state is equal to x;
- for all $k \in \mathbb{N}$, the set $\mathcal{A}_k := \{(x,u) \mid x \in \mathcal{E},\ u \in \mathcal{C}_k(x)\}$ of all possibles couples (state, action) at time k;
- an initial probability $p^{(0)} \in \Delta_{\mathcal{E}}$ on $\mathcal{E}$, or an initial state $x_0 \in \mathcal{E}$, which is equivalent to the case where $p^{(0)}$ is the Dirac measure at x_0;
- a probability space $(\Omega, \mathfrak{A}, P)$, a random variable X_0 with values in $\mathcal{E}$ and law $p^{(0)}$, and a sequence of independent random variables $(W_n)_{n\geq 0}$ with values in some discrete space $\mathcal{W}$, independent from X_0;
- for all $k \geq 0$, the *dynamics* at time k, which is a map $f_k : \mathcal{A}_k \times \mathcal{W} \to \mathcal{E}$.

The MDP is *stationary* if $\mathcal{C}_k(x)$ and f_k do not depend on time k, and if the W_k are identically distributed. In this case, the index or argument k is omitted. It is *uncontrolled* if the sets $\mathcal{C}_k(x)$ are singletons. In this case, the argument u is omitted.

Given a MDP in the sense of Definition 3.8, and a strategy of one of the above form, one can construct on a probability space (extending $(\Omega, \mathfrak{A}, P)$), two discrete time processes $(X_k)_{k\geq 0}$ and $(U_k)_{k\geq 0}$ taking their values in $\mathcal{E}$ and $\mathcal{C}$ respectively, satisfying:

$$X_{n+1} = f_n(X_n, U_n, W_n), \quad n \geq 0 \ . \tag{3.2}$$

Fact 3.9. Given a MDP in the sense of Definition 3.8, we can construct the following transition probabilities which define a MDP in the sense of Definition 3.1, with same behavior as the initial MDP:

$$M_{xy}^{(k,u)} = P(f_k(x,u,W_k) = y) \ .$$

Proof. Indeed, given the probability space as in Definition 3.8, and the random variables X_0 and W_n, for all pure strategies σ, one can associate the random process $(X_k, U_k)_{k\geq 0}$ satisfying both (3.2) and

$$U_k = \sigma_k(X_0, U_0, \ldots, X_{k-1}, U_{k-1}, X_k) \ .$$

In that case,

$$\begin{aligned} M_{xy}^{(k,u)} &= P(X_{k+1} = y \mid X_k = x, U_k = u) \\ &= P(f_k(x,u,W_k) = y) \ . \end{aligned}$$

Moreover, if $\mathcal{W}$ is finite or discrete, then

$$P(f_k(x,u,W_k) = y) = \sum_{w \in \mathcal{W},\ \text{s.t.}\ f_k(x,u,w)=y} P(W_k = w) \ .$$

If one consider a relaxed strategy however, one need to increase the probability space in order to handle all the possible probability laws $\sigma_k(X_0, U_0, \ldots, X_{k-1}, U_{k-1}, X_k)$. □

Associated to a Markov decision process, we can consider a *Markov decision problem* or a *discrete time stochastic control problem* which consists in maximizing (or minimizing) a criteria equal to the expected value of a functional of the random processes $(X_k)_{k\geq 0}$ and $(U_k)_{k\geq 0}$ induced by the above model among all (relaxed) strategies or among a restricted set of strategies. As for deterministic control problems, the criteria can be of several types:

- Finite horizon (time) additive or multiplicative or mixed criteria.
- Infinite horizon discounted (additive) criteria.
- Additive criteria with stopping time, which may be fixed or to be optimized.
- Long run time average criteria.

3.2 Markov decision problems with additive finite horizon criteria

Let be given a Markov decision process as in Definition 3.1 or Definition 3.8, and consider or denote:

- for all $k \in \mathbb{N}$, the *instantaneous/running reward/payoff* at time k, which is a map $r_k : \mathcal{A}_k \to \mathbb{R}$;
- a *final reward*, which is a map $\varphi : \mathcal{E} \to \mathbb{R}$;
- for all strategies $\sigma = (\sigma_k)_{k\geq 0}$ in Σ or Σ^{R} (or Π and Π^{R}), the *total additive payoff* with finite horizon $T \geq 1$:

$$J^{(T,\sigma)} := J^T(X;U) := \mathbb{E}\left[\left(\sum_{k=0}^{T-1} r_k(X_k,U_k)\right) + \varphi(X_T)\right] \quad , \tag{3.3}$$

  where $(X,U) := (X_k,U_k)_{k\geq 0}$ is the process induced by σ as in Definition 3.2 or Definition 3.3, on a probability space $(\Omega,\mathfrak{A},P)$.

- and for strategies associated to the MDP starting at time t, the *additive payoff starting at time t*:

$$J_t^{(T,\sigma)}(x) := J_{t,x}^T(X;U) := \mathbb{E}\left[\left(\sum_{k=t}^{T-1} r_k(X_k,U_k)\right) + \varphi(X_T) \mid X_t = x\right] \quad . \tag{3.4}$$

- for all $k \geq 0$ and $\pi \in \Pi_k$ (feedback policy at time k), the *associated reward vector at time k* $r^{(k,\pi)}$ with $r_x^{(k,\pi)} := r_k(x,\pi(x))$, for $x \in \mathcal{E}$.

Definition 3.10. A Markov decision problem with complete observation, and the above data consists in the following optimization problem:

$$\max_{\sigma} J^{(T,\sigma)}$$

where the optimization holds over either all relaxed strategies $\sigma \in \Sigma^{\mathrm{R}}$, or all pure strategies, or all Markov strategies, or all feedback policies. The optimum of above criteria is called the *value* of the problem. An optimal solution σ is called an *optimal strategy*, and the corresponding process U_k or (X_k,U_k) an *optimal control process*. Moreover, maximization can be replaced by minimization.

Definition 3.11. For all $x \in \mathcal{E}$, and $t \leq T$, let $v_t(x)$ be the value of the Markov decision problem with a criteria starting at time t:

$$\max_{\sigma} J_t^{(T,\sigma)}(x) \ .$$

The map $v : \{0, \ldots, T\} \times \mathcal{E} \to \mathbb{R}, (t, x \mapsto v_t(x)$ is called the *value function* of the MDP.

Using Theorem 2.18 and Fact 3.7, we deduce the following result.

Lemma 3.12. *When $\sigma = \pi = (\pi_k)_{k\geq 0}$ is a feedback policy, then $v_t := J_t^{(T,\pi)}$ satisfies the Kolmogorov equation:*

$$v_k = r^{(k,\pi_k)} + M^{(k,\pi_k)} v_{k+1},\ 0 \leq k \leq T-1 \ ,$$

with final condition:

$$v_T = \varphi \ .$$

In the controlled case, we obtain the nonlinear Bellman equation.

Theorem 3.13 (Dynamic programming equation for Markov decision problems with finite horizon)**.** *Assume that the maps φ, r_k, $k \geq 0$ are bounded from above. Let v_k be the value function of the Markov decision problem:*

$$v_k(x) := \max_{\sigma} J_k^{(T,\sigma)}(x) \ ,$$

where the maximum is taken over all relaxed strategies starting at time k. Then, v satisfies the following backward recurrence, called the Bellman dynamic programming equation*:*

$$v_k(x) = \sup_{u \in \mathcal{C}_k(x)} \left(r_k(x,u) + \sum_{y \in \mathcal{E}} M_{xy}^{(k,u)} v_{k+1}(y) \right) \quad \forall x \in \mathcal{E},\ 0 \leq k \leq T-1 \ . \tag{3.5}$$

with final condition

$$v_T = \varphi \ .$$

Moreover, the values v obtained by optimizing over the restricted sets of pure strategies, Markov strategies, or feedback policies, coincide.

Assume in addition that the maximum of (3.5) *is attained for an action $u \in \mathcal{C}_k(x)$ and* let us denote by $\pi_k(x)$ this action*, then the feedback policy $\pi = (\pi_k)_{0\leq k\leq T-1}$ is an optimal strategy of the problem.*

Note that we also have $v = p^{(0)} v_0$ for the value of the Markov decision problem with total additive payoff.

3.3 Properties of Bellman operators

We shall use similar operator notations as for deterministic control problems. For all $k \in \mathbb{N}$ and $\pi \in \Pi_k$, $\mathcal{B}^{(k,\pi)}$ and $\mathcal{B}^{(k)}$ denote the maps from $\mathbb{R}^{\mathcal{E}}$ to itself such that (assuming that the r_k are bounded from above)

$$\mathcal{B}^{(k,\pi)}(v) = r^{(k,\pi)} + M^{(k,\pi)} v$$

$$[\mathcal{B}^{(k)}(v)](x) = \sup_{u \in \mathcal{C}_k(x)} \left(r_k(x,u) + \sum_{y \in \mathcal{E}} M_{xy}^{(k,u)} v(y) \right) \ . \tag{3.6}$$

The dynamic programming equation (3.5) can be rewritten as $v_k(x) = [\mathcal{B}^{(k)}(v_{k+1})](x)$ for all $x \in \mathcal{E}$, or simply

$$v_k = \mathcal{B}^{(k)}(v_{k+1}). \tag{3.7}$$

Moreover, the supremum in (3.6) is finite as soon as $\mathcal{E}$ is finite, $v \in \mathbb{R}^{\mathcal{E}}$, and the functions r_k are bounded from above. Then, the operator $\mathcal{B}^{(k)}$ is well defined from $\mathbb{R}^{\mathcal{E}}$ to itself.

Remark 3.14. As for deterministic or uncontrolled problems, when the MDP is stationary, that is $\mathcal{C}_k(x)$ and $M_x^{(k,u)}$ do not depend on k, and the reward r_k is independent of time k too, then $\mathcal{B}$ does not depend on k, and one can consider the value function as a function of the remaining time until the end:

$$v^{(t)}(x) = \max_{\sigma} J_0^{(t,\sigma)}(x) \ .$$

Then, the backward Bellman equation for the value function is equivalent to the following *forward Bellman* equation:

$$v^{(k+1)}(x) = \sup_{u \in \mathcal{C}(x)} \left(r(x,u) + \sum_{y \in \mathcal{E}} M_{xy}^{(u)} v^{(k)}(y) \right) \quad \forall x \in \mathcal{E}, 0 \leq k \leq T-1 \ ,$$

or equivalently to

$$v^{(k+1)} = \mathcal{B}(v^{(k)}) \ ,$$

with initial condition $v^{(0)} = \varphi$.

Definition 3.15. For all $k \leq T-1$, the map $\mathcal{B}^{(k)}$ is called the *Bellman operator* at time k of the Markov decision problem.

The maps $\mathcal{B}^{(k,\pi)}$ are the *Kolmogorov or Bellman operators* at time k of the Markov decision problem, when the policy is freezed.

Since

$$[\mathcal{B}^{(k,\pi)}(v)](x) = r_k(x, \pi(x)) + \sum_{y \in \mathcal{E}} M_{xy}^{(k,\pi(x))} v(y)$$

only depends on $\pi(x)$, and not on all the other values of the map π, we get:

$$[\mathcal{B}^{(k)}(v)](x) = \sup_{\pi \in \Pi_k} [\mathcal{B}^{(k,\pi)}(v)](x) \quad \forall v \in \mathbb{R}^{\mathcal{E}}, \ \forall x \in \mathcal{E}.$$

which implies that

$$\mathcal{B}^{(k)}(v) = \sup_{\pi \in \Pi_k} \mathcal{B}^{(k,\pi)}(v) \ \forall v \in \mathbb{R}^{\mathcal{E}}, \tag{3.8}$$

where the suppremum is taken for the partial order of $\mathbb{R}^{\mathcal{E}}$.

Moreover, the existence of an optimum u in the dynamic programming equation (3.5) for all $x \in \mathcal{E}$, is equivalent to the existence of $\pi \in \Pi_k$ such that $\mathcal{B}^{(k)}(v_{k+1}) = \mathcal{B}^{(k,\pi)}(v_{k+1})$, that is to the property that the supremum in (3.8) is a maximum when $v = v_{k+1}$.

The operators $\mathcal{B}^{(k,\pi)} : \mathbb{R}^{\mathcal{E}} \to \mathbb{R}^{\mathcal{E}}$ are affine operators that are the Kolmogorov operators associated to the Markov chain of Fact 3.7. We already know that the operators $\mathcal{B}^{(k,\pi)}$ are monotone additively homogeneous (since a Markov matrix M has nonnegative entries and satisfies $M\mathbf{1} = \mathbf{1}$). The Bellman operators as suprema of such operators satisfy the same property:

Lemma 3.16. *The Bellman operators $\mathcal{B}^{(k)}$ are monotone and additively homogeneous.*

However, the Bellman operators $\mathcal{B}^{(k)}$ are in general nonlinear both in usual algebra and tropical algebra. They are however convex in the following sense, as suprema of affine maps.

Definition 3.17. An operator $\mathcal{B} : \mathbb{R}^{\mathcal{E}} \to \mathbb{R}^{\mathcal{E}}$ is *convex* if for all $x \in \mathcal{E}$, the function $v \in \mathbb{R}^{\mathcal{E}} \mapsto [\mathcal{B}(v)](x) \in \mathbb{R}$ is convex, or equivalently, for all $v, w \in \mathbb{R}^{\mathcal{E}}$ and $t \in [0,1]$, $\mathcal{B}((1-t)v + tw) \leq (1-t)\mathcal{B}(v) + t\mathcal{B}(w)$ for the partial order of $\mathbb{R}^S$.

Fact 3.18. The Bellman operators $\mathcal{B}^{(k)}$ are convex.

3.4 Proof of Theorem 3.13

We shall show the results by using the previous notations and properties, although one may do everything without using these properties explicitely.

Denote by $(w_k)_{0 \leq k \leq T}$ the sequence defined (uniquely) by the final condition $w_T = \varphi$, and the backward recurrence equations (3.5) or equivalently (3.7): $w_k = \mathcal{B}^{(k)}(w_{k+1})$. We denote also $w := p^{(0)} w_0$. We need to show that $w_k = v_k$ for all $k \in \{0, \ldots, T\}$ and that $w = v$.

1. Proof of $w_k \geq v_k$ and $w \geq v$ (without any assumption). Since for v_k, we use strategies starting at time k, the inequality $w_k \geq v_k$ is of same type as $w_0 \geq v_0$. So we show $w_0 \geq v_0$ only.

Let $\sigma \in \Sigma^{\mathrm{R}}$ be fixed, let $(X_t, U_t)_{t \geq 0}$ be the process associated to the strategy σ, and let $H_k = (X_0, U_0, X_1, U_1, \ldots, X_k)$, for all $k \geq 0$, be the history process. For all $0 \leq k \leq T$, and all histories $h_k = (x_0, u_0, \ldots, x_{k-1}, u_{k-1}, x_k) \in \mathcal{H}_k$, denote

$$z_k^{(\sigma)}(h_k) = \mathbb{E}\left[\left(\sum_{t=k}^{T-1} r_t(X_t, U_t)\right) + \varphi(X_T) \mid H_k = h_k\right] \quad .$$

Then, $z_0^{(\sigma)}(x_0) = J_0^{(T,\sigma)}(x_0)$, and so $J^{(T,\sigma)} = \mathbb{E}\left[z_0^{(\sigma)}(X_0)\right] = p^{(0)} z_0^{(\sigma)}$.

The process H_k is a Markov chain taking its values in the variable state space $\mathcal{H}_k$. Therefore, $z_k^{(\sigma)}$ satisfies the Kolmogorov equation for an additive functional (see Theorem 2.18)

$$z_n^{(\sigma)}(h_n) = \mathbb{E}\left[r_n(x_n, U_n) + z_{n+1}^{(\sigma)}(h_n, U_n, X_{n+1}) \mid H_n = h_n\right] \tag{3.9}$$

if $h_n = (x_0, u_0, \ldots, x_n) \in \mathcal{H}_n$, with the final condition $z_T^{(\sigma)}(h_T) = \varphi(x_T)$.

w_k satisfies (3.5), which can be rewritten as

$$w_n(x) = \sup_{u \in \mathcal{C}_n(x)} \left(r_n(x,u) + \mathbb{E}\left[w_{n+1}(X_{n+1}) \mid X_n = x, U_n = u\right]\right) \quad \forall x \in \mathcal{E} \quad . \tag{3.10}$$

Let us show that $w_k(x_k) \geq z_k^{(\sigma)}(h_k)$ by backward induction on k, when $h_k = (x_0, u_0, \ldots, x_k) \in \mathcal{H}_k$. This is true for $k = T$ since $z_T^{(\sigma)}(h_T) = \varphi(x_T) = w_T(x_T)$. If the inequality is true for $k+1$, then from the above equations (3.9) and (3.10), we deduce

$$\begin{aligned} z_k^{(\sigma)}(h_k) &\leq \mathbb{E}\left[r_k(x_k, U_k) + w_{k+1}(X_{k+1}) \mid H_k = h_k\right] = \\ &\mathbb{E}\left[r_k(x_k, U_k) + \mathbb{E}\left[w_{k+1}(X_{k+1}) \mid X_k = x_k, U_k\right] \mid H_k = h_k\right] = \\ &\sum_{u_k \in \mathcal{C}_k(x_k)} \sigma_k(h_k)(u_k) \left\{r_k(x_k, u_k) + \mathbb{E}\left[w_{k+1}(X_{k+1}) \mid X_k = x_k, U_k = u_k\right]\right\} \quad , \end{aligned}$$

where the last equality holds when $\sigma(h_k)$ is a probability with a countable support (in $\mathcal{C}_k(x_k)$).

Using (3.10), we deduce $z_k^{(\sigma)}(h_k) \leq \mathbb{E}\left[w_k(x_k) \mid H_k = h_k\right] = w_k(x_k)$, which proves the induction. In particular $z_0^{(\sigma)}(x_0) \leq w_0(x_0)$ for all $x_0 \in \mathcal{E}$. Since $z_0^{(\sigma)}(x_0) = J_0^{(T,\sigma)}(x_0)$, taking the maximum over all strategies, we deduce that $v_0(x_0) \leq w_0(x_0)$. Since $J^{(T,\sigma)} = p^{(0)} z_0^{(\sigma)} \leq p^{(0)} w_0$, taking again the maximum over all strategies, we deduce $v \leq w$.

2. Proof of $w_k \leq v_k$ and $w \leq v$ when the supremum in (3.5) is attained. Assume that the maximum in

$$w_k = \sup_{\pi \in \Pi_k} \mathcal{B}^{(k,\pi)}(w_{k+1}) \tag{3.11}$$

is attained for some policy $\pi \in \Pi_k$ (for instance if the sets $\mathcal{C}_k(x)$ are finite). Denote $\pi = (\pi_k)_{k \geq 0}$. Then,

$$w_k = \mathcal{B}^{(k,\pi_k)}(w_{k+1}), \quad k = 0, \ldots, T-1,$$

which means that $(w_k)_{n \geq 0}$ satisfies the same Kolmogorov equation as $J_k^{(T,\pi)}$, given in Lemma 3.12, with same final condition $w_T = \varphi = J_T^{(T,\pi)}$. Hence, $w_k = J_k^{(T,\pi)}$, for all $k \geq 0$. So $w_k \leq v_k$, where the value v_k is obtained as the maximum over any set of strategies containing at least feedback policies. Similarly, $w = p^{(0)} w^0 = J^{(T,\pi)} \leq v$. Moreover, π is an optimal strategy which is a feedback strategy.

3. Proof of $w_k \leq v_k$ and $w \leq v$ in general. Assume now that the maximum in (3.11) is not attained. We only assume that the maps r_k are bounded from above, for all $k \leq T-1$. This condition ensures in particular that the operators $\mathcal{B}^{(k)}$ are well defined as operators from $\mathbb{R}^{\mathcal{E}}$ to itself. The supremum in (3.6) is finite, therefore for all $\varepsilon > 0$, $k \leq T-1$ and $v \in \mathbb{R}^{\mathcal{E}}$, there exists $\pi \in \Pi_k$ such that

$$[\mathcal{B}^{(k,\pi)}(v)](x) \geq [\mathcal{B}^{(k)}(v)](x) - \varepsilon \quad \forall x \in \mathcal{E} \ ,$$

which can be rewritten as

$$\mathcal{B}^{(k,\pi)}(v) \geq \mathcal{B}^{(k)}(v) - \varepsilon \mathbf{1}.$$

Let $\pi_k \in \Pi_k$ such that

$$\mathcal{B}^{(k,\pi_k)}(w_{k+1}) \geq \mathcal{B}^{(k)}(w_{k+1}) - \varepsilon \mathbf{1}.$$

We have

$$\mathcal{B}^{(k,\pi_k)}(w_{k+1}) \geq w_k - \varepsilon \mathbf{1}.$$

Denote $z_k = w_k + (k-T)\varepsilon \mathbf{1}$. We have $z_T = w_T = \varphi$ and

$$\mathcal{B}^{(k,\pi_k)}(z_{k+1}) = \mathcal{B}^{(k,\pi_k)}(w_{k+1}) + (k+1-T)\varepsilon \mathbf{1} \geq w_k + (k-T)\varepsilon \mathbf{1} = z_k \tag{3.12}$$

since $\mathcal{B}^{(k,\pi_k)}$ is additively homogeneous.

Then, the functions z_k are sub-solutions of the Kolmogorov equation of Lemma 3.12. We shall show the inequality $z_k \leq J_k^{(T,\pi)}$ by backward induction on k. Indeed the inequality is true for $k = T$, since $z_T = \varphi = J_T^{(T,\pi)}$. If it holds for $k+1$, then

$$z_k \leq \mathcal{B}^{(k,\pi_k)}(z_{k+1}) \leq \mathcal{B}^{(k,\pi_k)}(J_{k+1}^{(T,\pi)}) = J_k^{(T,\pi)} \leq v_k \ ,$$

where the first inequality follows from (3.12), the second one from the induction assumption and the monotonicity of $\mathcal{B}^{(k,\pi_k)}$, the third one from the Kolmogorov equation of Lemma 3.12 which is satisfied by $J_k^{(T,\pi)}$, and the last one by definition of v_k. This shows $z_0 \leq J_0^{(T,\pi)} \leq v_0$, hence we get $p^{(0)} z_0 \leq p^{(0)} J_0^{(T,\pi)} = \mathbb{E}\left[J^{(T,\pi)}\right] \leq v$.

Therefore,

$$w_k = z_k + (T-k)\varepsilon \mathbf{1} \leq v_k + (T-k)\varepsilon \mathbf{1}$$

for all $k \in \{0, \dots, T\}$. Since we have shown this inequality for all $\varepsilon > 0$, we deduce that $w_k \leq v_k$ for all $k \in \{0, \dots, T\}$. Similarly $w = p^{(0)} w_0 = p^{(0)} z_0 + T\varepsilon \leq v + T\varepsilon$, and since this holds for all $\varepsilon > 0$, we get $w \leq v$. □

Remark 3.19. In the present case of an additive functional, another way to prove that the values v obtained by optimizing over either all relaxed strategies or all Markov strategies coincide is to show that for all $\pi \in \Sigma^{\mathrm{R}}$, there exists $\pi' \in \Pi^{\mathrm{R}}$ such that $\mathbb{E}\left[J^{(\pi)}\right] = \mathbb{E}\left[J^{(\pi')}\right]$. Indeed, let $(\pi'_k(x))(B) = P(U_k \in B, X_k = x)/P(X_k = x)$, where P is the probability on the process (X_k, U_k) induced by π as in Definition 3.3. Then, if (X'_k, U'_k) is the process induced by π', we get that for all $k \geq 0$, the laws of the random variables (X_k, U_k) and $X'_k, U'_k)$ coincide (not the ones of the processes). Therefore $\mathbb{E}\left[J^{(\pi)}\right] = \mathbb{E}\left[J^{(\pi')}\right]$, since they are both sums of expectations of functions of the random variables (X_k, U_k) only and not of all the process $(X_k, U_k)_{k\geq 0}$.

3.5 Problems with multiplicative or discounted finite horizon payoff

Let be given a Markov decision process as in Definition 3.1 or Definition 3.8, and consider or denote:

- for all $k \in \mathbb{N}$, the *instantaneous/running reward/payoff* at time k, which is a map $r_k : \mathcal{A}_k \to \mathbb{R}$;
- for all $k \in \mathbb{N}$, a variable *discount factor* at time k, which is a map $\alpha_k : \mathcal{A}_k \to \mathbb{R}_+$;
- a *final reward*, which is a map $\varphi : \mathcal{E} \to \mathbb{R}$;
- for all strategies $\sigma = (\sigma_k)_{k\geq 0}$ in Σ or Σ^{R}, the *mixed payoff* with finite horizon $T \geq 1$:

$$J^{(T,\sigma)} := J^T(X;U) := \mathbb{E}\left[\left(\sum_{\ell=0}^{T-1}\left(\prod_{m=0}^{\ell-1}\alpha_m(X_m,U_m)\right) r_\ell(X_\ell,U_\ell)\right) + \left(\prod_{m=0}^{T-1}\alpha_m(X_m,U_m)\right)\varphi(X_T)\right], \tag{3.13}$$

  where $(X,U) := (X_k,U_k)_{k\geq 0}$ is the process induced by σ as in Definition 3.2 or Definition 3.3.

- and for strategies associated to the MDP starting at time t, the *mixed payoff starting at time* t:

$$J_t^{(T,\sigma)}(x) := J_{t,x}^T(X;U) := \mathbb{E}\left[\left(\sum_{\ell=t}^{T-1}\left(\prod_{m=t}^{\ell-1}\alpha_m(X_m,U_m)\right) r_\ell(X_\ell,U_\ell)\right) + \left(\prod_{m=t}^{T-1}\alpha_m(X_m,U_m)\right)\varphi(X_T) \mid X_t = x\right]. \tag{3.14}$$

Theorem 3.20 (Dynamic programming equation for Markov decision problems with mixed functional and finite horizon)**.** *Assume that the maps φ, r_k, α_k, $k \geq 0$ are bounded from above. Let v_k be the value function of the Markov decision problem:*

$$v_k(x) := \max_{\sigma} J_k^{(T,\sigma)}(x) \ ,$$

where the maximum is taken over all relaxed strategies starting at time k. Then, v satisfies the following backward recurrence, called the Bellman dynamic programming equation*:*

$$v_k(x) = \sup_{u \in \mathcal{C}_k(x)} \left(r_k(x,u) + \alpha_k(x,u) \sum_{y \in \mathcal{E}} M_{xy}^{(k,u)} v_{k+1}(y) \right) \quad \forall x \in \mathcal{E},\ 0 \leq k \leq T-1 \ . \tag{3.15}$$

with final condition

$$v_T = \phi \ .$$

Moreover, the values v obtained by optimizing over the restricted sets of pure strategies, Markov strategies, or feedback policies, coincide.

Assume in addition that the maximum of (3.15) *is attained for an action $u \in \mathcal{C}_k(x)$ and* let us denote by $\pi_k(x)$ this action*, then the feedback policy $\pi = (\pi_k)_{0 \leq k \leq T-1}$ is an optimal strategy of the problem.*

Remark 3.21. As above, when the MDP is stationary, that is $\mathcal{C}_k(x)$ and $M_x^{(k,u)}$ do not depend on k, and the reward r_k and discount factor α_k are independent of time k too, then $\mathcal{B}$ does not depend on k, and one can consider the value function as a function of the remaining time until the end:

$$v^{(t)}(x) = \max_{\sigma} J_0^{(t,\sigma)}(x) \ .$$

Then, the backward Bellman equation for the value function is equivalent to the following *forward Bellman* equation:

$$v^{(k+1)}(x) = \sup_{u \in \mathcal{C}(x)} \left(r(x,u) + \alpha(x,u) \sum_{y \in \mathcal{E}} M_{xy}^{(u)} v^{(k)}(y) \right) \quad \forall x \in \mathcal{E}, 0 \leq k \leq T-1 \ ,$$

or equivalently to

$$v^{(k+1)} = \mathcal{B}(v^{(k)}) \ ,$$

with initial condition $v^{(0)} = \varphi$.

For the proof of Theorem 3.20, we shall use the Kolmogorov operators $\mathcal{B}^{(k,\pi)}$ from $\mathbb{R}^{\mathcal{E}}$ to itself defined, for $k \in \mathbb{N}$ and $v \in \mathbb{R}^{\mathcal{E}}$, by:

$$\mathcal{B}^{(k,\pi)}(v) = r_k^{(\pi)} + A^{(k,\pi)} v$$

where $A^{(k,\pi)}$ is the matrix with nonnegative entries such that $A_{xy}^{(k,\pi)} = \alpha_k(x, \pi(x)) M_{xy}^{(k,\pi)}$. We also use the Bellman operators $\mathcal{B}^{(k)}$ defined by (3.8). We have:

$$[\mathcal{B}^{(k)}(v)](x) = \sup_{u \in \mathcal{C}_k(x)} \left(r_k(x,u) + \alpha_k(x,u) \sum_{y \in \mathcal{E}} M_{xy}^{(k,u)} v(y) \right) \ . \tag{3.16}$$

Then, the dynamic programming equation (3.15) can be rewritten as in (3.7). As in the additive case, the supremum in (3.16) is finite as soon as $\mathcal{E}$ is finite, $v \in \mathbb{R}^{\mathcal{E}}$, and the functions $r_k(x,\cdot)$ and $\alpha_k(x,\cdot)$ are bounded from above, which implies that $\mathcal{B}^{(k)}$ is well defined and is a map from $\mathbb{R}^{\mathcal{E}}$ to itself, and that all the value functions v_k take finite real values. Moreover, the existence of an optimal control u in (3.15) for all $x \in \mathcal{E}$ is equivalent to the existence of a policy $\pi \in \Pi_k$ such that $\mathcal{B}^{(k)}(v_{k+1}) = \mathcal{B}^{(k,\pi)}(v_{k+1})$. The operators $\mathcal{B}^{(k,\pi)} : \mathbb{R}^{\mathcal{E}} \to \mathbb{R}^{\mathcal{E}}$ are affine, monotone, and thus the operators $\mathcal{B}^{(k)}$ are convex and monotone. However, they are no more additively homogeneous. Nevertheless, when $\alpha_k(x,u) \leq \beta$, for all $x \in \mathcal{E}$ and $u \in \mathcal{C}_k(x)$, the operators $\mathcal{B}^{(k,\pi)}$ and $\mathcal{B}$ are β-subhomogeneous, where this property is defined as follows.

Definition 3.22. Let $\beta > 0$. We say that an operator $\mathcal{B} : \mathbb{R}^{\mathcal{E}} \to \mathbb{R}^{\mathcal{E}}$ is *additively β-subhomogeneous* if it satisfies, for all $v \in \mathbb{R}^{\mathcal{E}}$ and $\lambda \in \mathbb{R}_+$,

$$\mathcal{B}(v + \lambda \mathbf{1}) \leq \mathcal{B}(v) + \beta\lambda \mathbf{1} \ .$$

When $\beta = 1$, we say that $\mathcal{B}$ is *additively subhomogeneous*

Proof of Theorem 3.20. We follow the same arguments as in the proof of Theorem 3.13, where Equation (3.15) is substituted to (3.5), and Kolmogorov equation of Theorem 2.23 is used instead of the one of Theorem 2.18. Indeed, Points 1 and 2 of the proof of Theorem 3.13 only use the monotonicity of the operators $\mathcal{B}^{(k,\pi)}$ and $\mathcal{B}^{(k)}$, Property (3.8) defining $\mathcal{B}^{(k)}$ as a supremum of the $\mathcal{B}^{(k,\pi)}$, and the Kolmogorov equation satisfied by the Markov chain H_k associated to any strategy, or the Markov chain X_k associated to any feedback policy. They thus can be followed similarly for the operators $\mathcal{B}^{(k,\pi)}$ and $\mathcal{B}^{(k)}$ of this section.

For Point 3, we use the property that the functions α_k are bounded from above. Let β be an upper bound. Then, from the above remarks, the operators $\mathcal{B}^{(k,\pi)}$ and $\mathcal{B}^{(k)}$ are additively β-subhomogeneous. Choose π_k as in Point 3 of the proof of Theorem 3.13, and $z_k = w_k - \varepsilon\gamma_k$, where the sequence γ_k is obtained by the backward induction $\gamma_T = 0$, and $\gamma_k = 1 + \beta\gamma_{k+1}$ (so $\gamma_k = T - k$ if $\beta = 1$, and $\gamma_k = (\beta^{T-k} - 1)/(\beta - 1)$ otherwise). We obtain in the same way as in the proof of Theorem 3.13 that $\mathcal{B}^{(k,\pi_k)}(z_{k+1}) \geq z_k$, and thus $z_k \leq v_k^{(\pi)}$, and $p^{(0)} z_0 \leq v$. Therefore

$$w_k = z_k + \gamma_k\varepsilon \leq v_k^{(\pi)} + \gamma_k\varepsilon \leq v_k + \gamma_k\varepsilon$$

for all $k \in \{0, \dots, T\}$. Since this holds for all $\varepsilon > 0$, we deduce that $w_k \leq v_k$ for all $k \in \{0, \dots, T\}$. The rest of the proof remains. □

Remark 3.23. Another way to prove Theorem 3.20 is to increase the state space by taking $\mathcal{E}' := S \times \mathcal{Z}$, with $\mathcal{Z} = \mathbb{R}^+$, and considering the state sequence $X'_k = (X_k, Z_k)$, where $Z_k \in \mathcal{Z}$ satisfies $Z_{k+1} = \alpha_k(X_k, U_k)Z_k$ starting at $Z_0 = 1$. This corresponds to a MDP with transition probabilities given by $M'^{(k,u)}_{x'y'} = M^{(k,u)}_{xy}$ if $x' = (x,z)$ and $y' = (y, \alpha_k(x,u)z)$ and $M^{(k,u)}_{x'y'} = 0$ otherwise, and the initial law $p'^{(0)} = p^{(0)} \otimes p_z$, where p_z is the Dirac probability at Point 1 on $\mathcal{Z}$. The only difficulty is that now the new state space is infinite not countable. One can reduce however the problem to a finite state space $\mathcal{E}'_k$ depending on time if we assume that the action spaces $\mathcal{C}_k(x)$ are finite. Consider the reward functions defined for $x' = (x,z) \in \mathcal{E} \times \mathcal{Z}$, by $r'_k(x',u) = zr_k(x,u)$, $\varphi'(x') = z\varphi(x)$, we get that $J^T_{t,x}(X;U)$ is equal to the additive functional $J^T_{t,(x,1)}(X';U)$ defined with the new MDP and rewards. We can then apply Theorem 3.13. We deduce that $v = (p^{(0)} \otimes p_z)(v'_0)$ where $v'_T = \varphi'$

and v'_k satisfies the dynamic programming equation

$$v'_k(x') = \sup_{u \in \mathcal{C}_k(x)} \left(r'_k(x', u) + \sum_{y' \in \mathcal{E}'} M'^{(k,u)}_{x'y'} v'_{k+1}(y') \right) \quad \forall x' \in \mathcal{E}' \ .$$

Replacing r'_k and φ' with their values, we obtain that, for all $k \geq 0$, $v'_k(x') = z v_k(x)$ for some function v_k, and that v_k satisfies (3.15). Similarly $v = p'^{(0)} v_0$.

The previous remark shows that one can always reduce a MDP with mixed functional to a MDP with additive functional by increasing the state space which may become infinite. One can also do the reverse operation, when the state has the form $X_k = (X'_k, Z_k)$, in which Z_k is positively homogeneous, that is satisfies that Z_k is transformed into λZ_k when Z_0 is transformed into λZ_0, if X'_{k+1} does not depend on Z_k. Indeed, in that case the state can be reduced to X'_k, then the additive functional becomes a mixed functional.

When the discount factors α_k are less than 1, we call the mixed functional a *discounted functional*. Another reduction can be obtained in that case, by adding a cemetery point, as in Proposition 2.24. This leads to the following result.

Proposition 3.24. *The value function of a finite horizon discounted Markov Decision problem is equal to the restriction to $\mathcal{E}$ of the value function of a finite horizon MDP with additive criteria on the state space $\mathcal{E} \cup \{c\}$, where $c \notin \mathcal{E}$ is a* cemetery point*.*

Proof. Consider a MDP with finite horizon mixed functional as above. Let $c \notin \mathcal{E}$, and consider $\mathcal{E}' = \mathcal{E} \cup \{c\}$. We construct a MDP with additive functional on $\mathcal{E}'$ as follows. We keep the same action spaces $\mathcal{C}_k(x)$ for $x \in \mathcal{E}$ and we take for $\mathcal{C}_k(c)$ any singleton subset of $\mathcal{C}$, The initial law $p'^{(0)}$ is $p'^{(0)}_x = p^{(0)}_x$ for $x \in \mathcal{E}$ and $p'^{(0)}_c = 0$, and the transition probabilities are

$$\begin{aligned}
M'^{(k,u)}_{xy} &= \alpha_k(x,u) M^{(k,u)}_{xy}, \quad \text{when } x, y \in \mathcal{E} \\
M'^{(k,u)}_{xc} &= 1 - \alpha_k(x,u), \quad \forall x \in \mathcal{E} \\
M'^{(k,u)}_{cy} &= 0, \quad \forall y \in \mathcal{E} \\
M'^{(k,u)}_{cc} &= 1 \ .
\end{aligned}$$

We extend r_k and φ to $\mathcal{E}'$ in r'_k and φ' respectively by mapping (c,u) or c to 0. Let v' be the value function of this new MDP on $\mathcal{E}'$ with additive functional (see (3.4)). By Theorem 3.13, we have $v' = p'^{(0)} v'_0$ where $v'_k \in \mathbb{R}^{\mathcal{E}'}$ is solution of the backward recurrence dynamic programming equation

$$v'_k(x) = \sup_{u \in \mathcal{C}_k(x)} \left(r'_k(x,u) + \sum_{y \in \mathcal{E}'} M'^{(k,u)}_{xy} v'_{k+1}(y) \right) \quad \forall x \in \mathcal{E}', \ 0 \leq k \leq T-1 \ ,$$

with final condition $v'_T = \varphi'$. Since $\varphi(c) = r_k(c,u) = 0$ and $M'^{(k,u)}_{cc} = 1$, for all $u \in \mathcal{C}_k(c)$, we get that $v'_k(c) = v'_{k+1}(c)$ for all $k \geq 0$, hence $v'_k(c) = 0$. Therefore, if v_k is the restriction of v'_k to $\mathcal{E}$, we obtain that $v' = p^{(0)} v_0$ and that v_k satisfies the dynamic programming equation (3.15). Moreover, $v' = v$ so the values of the two problems coincide. □

3.6 Problems with exit time in finite horizon

Let be given a Markov decision process as in Definition 3.1 or Definition 3.8, and let B be a strict subset of $\mathcal{E}$. Then, any strategy $\sigma = (\sigma_n)_{n\geq 0}$ in Σ or Σ^{R} induces the process $(X_n, U_n)_{n\geq 0}$ and the history process $(H_n)_{n\geq 0}$. The later beeing a Markov chain, we can construct the filtration $(\mathcal{F}_n)_{n\geq 0}$ generated by $(H_n)_{n\geq 0}$, which is in that case $\mathcal{F}_n = \sigma^{\mathrm{a}}(H_n) = \sigma^{\mathrm{a}}(X_0, U_0, \ldots, U_{n-1}, X_n)$. Then, the exit time of the sequence $(X_n)_{n\geq k}$ from the set B:

$$\tau_B^k := \inf\{n \geq k \mid X_n \notin B\}$$

is a stopping time with respect to the filtration $(\mathcal{F}_n)_{n\geq k}$. When σ is a feedback policy, then $(X_n)_{n\geq 0}$ is a Markov chain and τ_B^k is also a stopping time with respect to the Markov chain $(X_n)_{n\geq k}$.

Consider or denote:

- for all $k \in \mathbb{N}$, the *instantaneous/running reward/payoff* at time k, which is a map $r_k : \mathcal{A}_k \to \mathbb{R}$;
- a *final reward*, which is a map $\varphi : \mathcal{E} \to \mathbb{R}$;
- for all $k \in \mathbb{N}$, an *exit reward*, which is a map $\psi_k : \mathcal{E} \setminus B \to \mathbb{R}$, such that ψ_T is the restriction of φ on $\mathcal{E} \setminus B$;
- for all strategies $\sigma = (\sigma_k)_{k\geq 0}$ in Σ or Σ^{R}, the *payoff with exit stopping time* and finite horizon $T \geq 1$:

$$J^{(T,B,\sigma)} := J^{T,B}(X;U) := \mathbb{E}\left[\left(\sum_{\ell=0}^{T\wedge\tau_B^0-1} r_\ell(X_\ell, U_\ell)\right) + \varphi(X_T)\mathbf{1}_{T<\tau_B^0} + \psi_{\tau_B^0}(X_{\tau_B^0})\mathbf{1}_{T\geq\tau_B^0}\right] , \tag{3.17}$$

  where $(X, U) := (X_k, U_k)_{k\geq 0}$ is the process induced by σ as in Definition 3.2 or Definition 3.3, and τ_B^0 is the exit time of the process $(X_n)_{n\geq 0}$ from B.

- and for strategies associated to the MDP starting at time t, the *payoff starting at time t with exit stopping time*:

$$J_t^{(T,B,\sigma)}(x) := J_{t,x}^{T,B}(X;U) := \mathbb{E}\left[\left(\sum_{\ell=t}^{T\wedge\tau_B^t-1} r_\ell(X_\ell, U_\ell)\right) + \psi_{T\wedge\tau_B^t}(X_{T\wedge\tau_B^t}) \mid X_t = x\right] , \tag{3.18}$$

  where we extend ψ_T to B by taking $\psi_T = \varphi$.

Theorem 3.25 (Dynamic programming equation for Markov decision problems with exit time in finite horizon)**.** *Assume that the maps φ, r_k, $k \geq 0$ are bounded from above. Let v_k be the value function of the Markov decision problem:*

$$v_k(x) := \max_\sigma J_k^{(T,B,\sigma)}(x) ,$$

where the maximum is taken over all relaxed strategies starting at time k. Then, v satisfies the following backward recurrence, called the Bellman dynamic programming equation*:*

$$v_k(x) = \sup_{u\in\mathcal{C}_k(x)} \left(r_k(x,u) + \sum_{y\in\mathcal{E}} M_{xy}^{(k,u)} v_{k+1}(y)\right) \quad \forall x \in B,\ 0 \leq k \leq T-1 . \tag{3.19a}$$

with boundary condition

$$v_k(x) = \psi_k(x), \quad \forall x \notin B,\ 0 \le k \le T \ , \tag{3.19b}$$

and final condition

$$v_T(x) = \varphi(x) = \psi_T(x), \quad \forall x \in B \ . \tag{3.19c}$$

Moreover, the values v obtained by optimizing over the restricted sets of pure strategies, Markov strategies, or feedback policies, coincide.

Assume in addition that the maximum of (3.19) *is attained for an action $u \in \mathcal{C}_k(x)$, for $x \in B$, and* let us denote by $\pi_k(x)$ this action *when $x \in B$, and choose any action $\pi_k(x)$ for $x \notin B$, then the feedback policy $\pi = (\pi_k)_{0 \le k \le T-1}$ is an optimal strategy of the problem.*

Theorem 3.25 can be deduced easily from Theorem 3.20 using the following property which is the same as Fact 2.34.

Fact 3.26. The functional $J_t^{(T,B,\sigma)}$ of (3.18) can rewritten as the mixed functional:

$$J_t^{(T,B,\sigma)}(x) := J_{t,x}^{T,B}(X;U) := \mathbb{E}\left[\left(\sum_{\ell=t}^{T-1}\left(\prod_{m=t}^{\ell-1}\alpha_m(X_m,U_m)\right) r'_\ell(X_\ell,U_\ell)\right) + \left(\prod_{m=t}^{T-1}\alpha_m(X_m,U_m)\right)\varphi(X_T) \mid X_t = x\right] \ .$$

for the same Markov Decision process, with the same final reward φ, and the instantaneous rewards r'_k and variable discount factors α_k given by:

$$\begin{aligned} r'_k(x,u) &= r_k(x,u), \quad \text{for } x \in B,\ u \in \mathcal{C}_k(x) \\ r'_k(x,u) &= \psi_k(x), \quad \text{for } x \notin B,\ u \in \mathcal{C}_k(x) \\ \alpha_k(x,u) &= 1, \quad \text{for } x \in B,\ u \in \mathcal{C}_k(x) \\ \alpha_k(x,u) &= 0, \quad \text{for } x \notin B,\ u \in \mathcal{C}_k(x) \ . \end{aligned}$$

The previous property shows that a Markov decision process with exit time in a finite horizon functional can be seen as a problem with mixed functional with discount factors ≤ 1, for the same MDP. By Proposition 3.24, such a problem can then be reduced to a problem with additive functional by adding a cemetery point to the state space. Another way to prove Theorem 3.25 or to reduce the exit time functional to an additive functional is to consider the process $X_{n\wedge\tau_B^k}$ which corresponds to a slightly different MDP, in which the transition probabilities in $\mathcal{E} \setminus B$ are changed (see the proof of Theorem 2.33). The advantage of Fact 3.26 is that we do not need to change the MDP.

Example 3.27. Conversely, consider a Markov decision process with an additive criteria as (3.3) including a cemetery point $c \in \mathcal{E}$, as in the reductions of previous section. This means that $M_{cc}^{(k,u)} = 1$, and $r_k(c,u) = 0$ for all $u \in \mathcal{C}_k(c)$. Then, one can rewrite the additive functional by using the exit stopping time τ_B from the set $B = \mathcal{E} \setminus \{c\}$ of the process $(X_k)_{k\ge 0}$ induced by any strategy σ. Indeed, in that case $r_k(X_k,U_k) = 0$ and $X_k = X_{k+1} = X_{\tau_B}$ for all $k \ge \tau_B$, and so

$$J^{(T,\sigma)} := J^T(X;U) := \mathbb{E}\left[\left(\sum_{k=0}^{T-1} r_k(X_k,U_k)\right) + \varphi(X_T)\right] = \mathbb{E}\left[\left(\sum_{k=0}^{T\wedge\tau_B-1} r_k(X_k,U_k)\right) + \varphi(X_{T\wedge\tau_B})\right] \ .$$

Again the advantage of this technique is that we do not need to change the MDP (the transition probabilities).

3.7 The example of optimal stopping time problems with finite horizon

Consider

- a (fixed) Markov chain $(X_n)_{n\geq 0}$ over a probability space $(\Omega, \mathfrak{A}, P)$ taking its values in a finite state space $\mathcal{E}$, with transition matrices $M^{(n)}$ at time $n \in \mathbb{N}$, and initial probability law $p^{(0)}$.
- *instantaneous/running rewards/payoffs* $r_k \in \mathbb{R}^B$ (at any time $k \leq T-1$);
- final rewards $\varphi_k \in \mathbb{R}^{\mathcal{E}}$ (at any time $k \leq T$);
- for all stopping times τ with respect to the Markov chain $(X_n)_{n\geq 0}$, the *finite horizon payoff with stopping time* τ:

$$J^{(T,\tau)} := \mathbb{E}\left[\left(\sum_{\ell=0}^{\tau\wedge T-1} r_\ell(X_\ell)\right) + \varphi_{\tau\wedge T}(X_{\tau\wedge T})\right] \ ; \tag{3.20}$$

- and for all $t \leq T$, the *finite horizon payoff with stopping time* $\tau \geq t$, starting in x at time t:

$$J_t^{(T,\tau)}(x) := \mathbb{E}\left[\left(\sum_{\ell=t}^{\tau\wedge T-1} r_\ell(X_\ell)\right) + \varphi_{\tau\wedge T}(X_{\tau\wedge T}) \mid X_t = x\right] \ . \tag{3.21}$$

Definition 3.28. An *Optimal stopping time problem with complete observation and finite horizon criteria* consists in the following optimization problem:

$$\max_\tau J^{(T,\tau)}$$

where the optimization holds over all stopping times τ with respect to the Markov chain $(X_n)_{n\geq 0}$.

The optimum of above criteria is called the *value* of the problem.

An optimal solution τ is called an *optimal stopping time*.

Definition 3.29. For all $x \in \mathcal{E}$, and $t \leq T$, let $v_t(x)$ be the value of the optimal stopping time problem with initial state x at time t:

$$\max_\tau J_t^{(T,\tau)}(x) \ .$$

The map $v : \{0, \ldots, T\} \times \mathcal{E} \to \mathbb{R}, (t, x) \mapsto v_t(x)$ is called the *value function* of the stopping time problem.

Theorem 3.30 (Dynamic programming equation for optimal stopping time problems with finite horizon criteria)**.** *Assume that the maps φ_k, r_k, $k \geq 0$, are bounded from above (or $\mathcal{E}$ finite). Let v_k be the value function of the optimal stopping time problem with finite horizon:*

$$v_k(x) := \max_\tau J_k^{(T,\tau)}(x) \ ,$$

where the maximum is taken over all stopping times τ with respect to the Markov chain $(X_n)_{n\geq k}$. Then, v satisfies the following backward recurrence, called the Bellman dynamic programming equation *or* variational inequality*:*

$$v_k(x) = \max\Big(r_k(x) + \sum_{y\in\mathcal{E}} M_{xy}^{(k)} v_{k+1}(y), \varphi_k(x)\Big) \quad \forall x \in \mathcal{E},\ 0 \leq k \leq T-1\ . \tag{3.22}$$

with final condition

$$v_T = \varphi_T\ .$$

For all $0 \leq k \leq T$, let B_k be the set of states in which the maximum in (3.22) is attained in the first term, that is

$$B_k := \{x \in \mathcal{E} \mid r_k(x) + \sum_{y\in\mathcal{E}} M_{xy}^{(k)} v_{k+1}(y) \geq \varphi_k(x)\}\ ,$$

au define

$$\tau = \inf\{k \geq 0 \mid X_k \notin B_k \text{ or } k = N\}\ .$$

Then τ is an optimal stopping time.

Proof. For the proof, we reduce this problem to a Markov decision problem with finite horizon criteria.

Consider the MDP in which

- the state space is $\mathcal{E}' = \mathcal{E} \cup \{c\}$ where $c \notin \mathcal{E}$;
- the control space $\mathcal{C} = \{0, 1\}$ (0 for stop, and 1 for not stop);
- the control space $\mathcal{C}(x)$ is such that $\mathcal{C}(x) = \mathcal{C}$ if $x \in \mathcal{E}$ and $\mathcal{C}(x) = \{0\}$ if $x = c$.
- the states of the MDP, Y_n, depend on the states of the Markov chain X_n and on the actions as follows:
$$Y_{n+1} = g(X_{n+1}, U_n)$$
where $g(x, u) = x$ if $u = 1$ and $g(x, u) = c$ otherwise.
- Then, for all $x_i \in \mathcal{E}',\ u_i \in \mathcal{C}_i(x_i)$, $i \geq 0$, we have
$$\begin{aligned} &P(Y_{k+1} = x_{k+1} \mid Y_k = x_k, U_k = u_k, Y_{k-1} = x_{k-1}, \ldots, Y_0 = x_0, U_0 = u_0) \\ &\quad = 1 \text{ if } c = x_{k+1} \text{ and } u_k = 0 \\ &\quad = 0 \text{ if } x_{k+1} \in \mathcal{E} \text{ and } u_k = 0 \\ &\quad = M_{x_k, x_{k+1}}^{(k)} \text{ if } x_k, x_{k+1} \in \mathcal{E} \text{ and } u_k = 1 \end{aligned}$$
- This implies that
$$\begin{aligned} &P(Y_{k+1} = x_{k+1} \mid Y_k = x_k, U_k = u_k, Y_{k-1} = x_{k-1}, \ldots, Y_0 = x_0, U_0 = u_0) \\ &\quad = P(Y_{k+1} = x_{k+1} \mid Y_k = x_k, U_k = u_k)\ . \end{aligned}$$
which is the Markov property.

- The transition probabilities of the MDP are: $M^{(k,1)}_{xy} = M^{(k)}_{xy}$ for all $x, y \in \mathcal{E}$, $M^{(k,1)}_{xy} = 0$ for $x \in \mathcal{E}$ and $y = c$, $M^{(k,0)}_{xy} = 1$ for $x \in \mathcal{E}'$ and $y = c$, and $M^{(k,0)}_{xy} = 0$ for $x \in \mathcal{E}'$ and $y \in \mathcal{E}$. Note that $M^{(k,1)}_{cy}$, with $y \in \mathcal{E}'$, is useless.

- Take then the rewards: $r'_k(x, 1) = r_k(x)$, $r'_k(x, 0) = \varphi_k(x)$ for $x \in \mathcal{E}$ and $r'_k(c, 0) = 0$, and final reward $\varphi(x) = \varphi_T(x)$ for $x \in \mathcal{E}$ and $\varphi(c) = 0$.

Then, the value of the finite horizon Markov Decision problem coincides with the value of the stopping time problem with finite horizon:

Given a pure strategy, the associated process U_t satisfies $U_t \in \mathcal{F}_t = \sigma^{\mathrm{a}}(X_0, \ldots, X_t)$, and so $\tau = \inf\{t \geq 0 \mid U_t = 0\}$ is a stopping time. Conversely, given a stopping time τ, consider the process such that $U_n = 1$ for all $n < \tau$ and $U_n = 0$ for $n \geq \tau$, we get that $\{U_t = 1\} \in \mathcal{F}_t = \sigma^{\mathrm{a}}(X_0, \ldots, X_t)$, so that U_t can be written as a measurable function σ_t of $X_0, \ldots, X_t$. Then, the process $(U_t)_{t\geq 0}$ is associated to the pure strategy $(\sigma_t)_{t\geq 0}$.

The Bellman equation of the Markov decision problem is then:

$$v_k(x) = \max\Big(r_k(x) + \sum_{y\in\mathcal{E}} M^{(k)}_{xy} v_{k+1}(y), \varphi_k(x) + v_{k+1}(c)\Big) \quad \forall x \in \mathcal{E} \ ,$$
$$v_k(c) = v_{k+1}(c) \ ,$$

with the final condition $v_T(x) = \varphi_T(x)$ for $x \in \mathcal{E}$ and $v_T(c) = 0$.

Here the action $u = 1$ corresponds to the left term in the above maximum, and $u = 0$ corresponds to the right term.

Therefore if $\pi_k : \mathcal{E} \to \{0, 1\}$ is an optimal policy given by the Bellman equation, we recover again the optimal stopping time by taking:

$$\tau = \inf\{t \geq 0 \mid \pi_t(X_t) = 0\}$$

or by taking τ as in the theorem. □

Remark 3.31. As seen in the above result, the variational inequality (3.22) allows one to construct the set $\cup_{k\geq 0}\{k\} \times B_k$, for which the optimal stopping time is the exit time of the process $(k, X_k)_{k\geq 0}$. The complementary of this set plays the same role as a *free boundary* in the continuous time and state setting. Contrarilly to the case of control problems with an exit time considered in Theorem 3.25, the above "boundary" is not known in advance, but is computed as the optimal boundary for a certain criterion.

3.8 Examples and Exercices

Example 3.32 (A random ressource allocation problem)**.** Let us consider the ressource allocation problem described in Example 1.7 and Example 1.15, where we assume now that the reward obtained when one invests x units in the ith ressource is random, and can be written as $R_i(x, Z_i)$, where the R_i are deteministic maps and the Z_i are independent random variables with values in some set $\mathcal{Z}$. Assume that the investor is choosing the numbers of units u_i he is investing in each ressource i in advance without knowing the values of the variables Z_i.

What is the maximal expected total reward of the investor ?

Assume now that the investor is choosing the amounts to be invested in the ressources sequentially, that is using some given order $\sigma(i), i = 1, \dots, N$, where σ is a permutation of $\{1, \dots, N\}$, and that when he decides to invest in ressource i, he is able to obtain the information on its reward, that is on Z_i. Assume also that the set $\mathcal{Z}$ is finite. The investor wants to maximize his expected total reward. Write this problem as a Markov decision problem with state variable (x, z) with $x \in \{0, \dots, R\}$ and $z \in \mathcal{Z}$.

Show that this problem can be solved via the recursive equation with final value $w_{N+1} = 0$.

$$w_n(x) = \mathbb{E}\left[\max_{u \in \mathbb{N},\, 0 \le u \le x} [r_{\sigma(n)}(u, Z_{\sigma(n)}) + w_{n+1}(x-u)]\right],\ n = 1, \dots, N, \quad x \in \{0, \dots, R\}\ ,$$

and that $w_1(R)$ gives the expected total reward of the investor. Show that if the maps r_i are concave with respect to the first variable, then the maps w_n are also concave (one can first consider the relaxed problem where x and the u_i can take real values).

Show that the solution depends on the permutation σ.

3.9 Problem: Airline Revenue Management

Consider an airline revenue manager who decide at each time of arrival of a demand of seats on a given flight, if he accept or reject this demand.

We assume that the flight contains $n > 1$ classes of seats (a class may contain an information of level of the seat, of position in the plane, and of date of booking as well), numbered from 1 to n. The price of a seat of class $k \in \{1, \dots, n\}$ will be denoted by p_k, and we have $p_1 < \cdots < p_n$. We assume also that demands of seats of class k arrive before demands of seats of class $k+1$, this means that demands arrive in nonoverlapping intervals in the order of increasing prices (this is called the early bird hypotheses). If D_k denotes the total amount of demands of seats of class k, we also assume that the D_k are independent random variables.

We consider a Markov decision problem with finite horizon (n stages), starting at stage 1, modelizing the sale of the seats of one flight of a plane only. At each stage $k = 1, \dots, n$, all the D_k demands of seats of class k arrive, and the manager decides to accept only a certain number of them $U_k \leq D_k$. We add a final stage $n+1$, at which no demand of seats is done, so $D_{n+1} = 0$. Moreover, for each $k = 1, \dots, n+1$, we denote by X_k the remaining capacity of the plane at stage k, that is the number of available seats. Thus, X_1 is the total number of seats of the plane, and X_{n+1} is the number of unsold seats at the end of the sale.

The Markov decision problem will involve a Markov decision process such that at each stage $k = 1, \dots, n+1$,

(X_k, D_k) : is the state of the MDP at stage k. Since the plane has a finite number of seats M, one can consider that the set of states $\mathcal{E}$ is of the form $\mathcal{E} = [M] \times \mathbb{N}$ where $[M] := \{0, \dots, M\}$, or even $\mathcal{E} = [M] \times [M]$.

U_k : is the action of the MDP at stage k. Again, since $U_k \leq D_k$, one can consider that the set of actions $\mathcal{C}$ is $[M]$, and that at stage k, U_k satisfies the constraint $0 \leq U_k \leq \min(X_k, D_k)$, so that $U_k \in \mathcal{C}(X_k, D_k)$, with $\mathcal{C}(x, d) = [\min(x, d)]$, for $(x, d) \in \mathcal{E}$.

Moreover, the dynamics of the MDP is such that:

$$X_{k+1} = X_k - U_k, \quad k = 1, \dots, n \ ,$$
$$D_k \text{ are independent random variables with given laws } q_k(d) = \mathbb{P}(D_k = d) \, , \ d \in \mathbb{N} \ .$$

The payoff (criteria) of the MDP is the expected amount of money obtained after the end of the sale, that the manager want to maximize. So the instantaneous reward at stage k, state (x, d) and action u is given by

$$r(k; x, d; u) = p_k u \ ,$$

and the final reward at state (x, d) is

$$\phi(x, d) = 0 \ .$$

Q 9.1. We denote by $v(k; x, d)$ the value function of the MDP, when the starting stage is $k = 1, \dots, n+1$ with starting state $(x, d) \in \mathcal{E}$:

$$v(k; x, d) = \sup \mathbb{E}\left[\sum_{\ell=k}^{n} r(\ell; X_\ell, D_\ell; U_\ell) + \phi(X_{n+1}, D_{n+1}) \mid X_k = x, D_k = d \right] \ ,$$

where the supremum is taken over all (feedback) strategies $(\pi_k)_{k \geq 0}$ defining the admissible process $(U_k)_{k \geq 1}$: $U_k = \pi_k(X_k, D_k)$. Write the dynamic programming equation satisfied by v.

Q 9.2. Denote $w_k(x) = \mathbb{E}\left[v(k; x, D_k)\right]$. Write a recurrence equation satisfied by w_k.

Q 9.3. Show by induction on k, that $w_k : [M] \to \mathbb{R}$ is concave, and that for each $d \in \mathbb{N}$, $x \in [M] \mapsto v(k; x, d) \in \mathbb{R}$ is concave.

Q 9.4. For $k = 2, \ldots, n+1$, choose

$$y_k \in \operatorname*{Argmax}_{y \in [M]} \left\{-p_{k-1}y + w_k(y)\right\} \quad .$$

Show that

$$\pi_k(x, d) = \max(\min(x - y_{k+1}, d), 0) = \min(\max(x - y_{k+1}, 0), d)$$

is an optimal policy at time k for our problem.

Q 9.5. Explain the meaning of y_k as a protection level for classes of levels $\geq k$.

Q 9.6. For all $k = 2, \ldots, n+1$, write y_k as a function of the map $\Delta w_k : [M] \to \mathbb{R} \cup \{+\infty\}$, where

$$\Delta w_k(x) = w_k(x) - w_k(x-1) \quad .$$

Q 9.7. To realize the previous policy, what should be the policy of acceptation of a single demand of a seat of class k?

Q 9.8. Compute y_n as a function of p_{n-1}, p_n and the law q_n of D_n.

Q 9.9. Show that $\Delta w_{k+1}(y) \leq \Delta w_k(y)$ for all $y \in \mathbb{N}$ and $k \in \{1, \ldots, n\}$.

Q 9.10. Show that $y_1 \geq \cdots \geq y_n$.

Some references for this problem:

[RM1] K. Littlewood. Forecasting and control of passenger bookings. In *Proc. 12th AGIFORS Symposium*. 1972. reprinted in Journal of Revenue and Pricing Management, Vol. 4 (2005).

[RM2] P.P. Belobaba. *Air Travel Demand and Airline Seat Inventory Management*. PhD thesis, Flight Transportation Laboratory. Cambridge, MIT, 1987.

[RM3] S.L. Brumelle and J.I. McGill. Airline seat allocation with multiple nested fare classes. *Operations Research*, (1):127–137, 1993.

[RM4] Kalyan T. Talluri and Garrett J. van Ryzin. *The theory and practice of revenue management*. International Series in Operations Research & Management Science, 68. Kluwer Academic Publishers, Boston, MA, 2004.

Chapter 4

Markov decision problems with infinite horizon

4.1 Discounted infinite horizon problems

Assume given a stationary Markov decision process, that is

- a finite or discrete *state space* $\mathcal{E}$;
- an *action space* $\mathcal{C}$
- for all $x \in \mathcal{E}$, the subset $\mathcal{C}(x) \subset \mathcal{C}$ of all possible actions at any time k, when the state is equal to x;
- the set $\mathcal{A} := \{(x, u) \mid x \in \mathcal{E},\ u \in \mathcal{C}(x)\}$ of all possibles couples (state, action) at any time k;
- an initial probability $p^{(0)} \in \Delta_{\mathcal{E}}$ on $\mathcal{E}$, or an initial state $x_0 \in \mathcal{E}$, which is equivalent to the case where $p^{(0)}$ is the Dirac measure at x_0;

with either

- for all $x \in \mathcal{E}$ and $u \in \mathcal{C}(x)$, a probability row vector $M_x^{(u)}$ over $\mathcal{E}$, the entries of which will be denoted $\left(M_{xy}^{(u)}\right)_{y \in \mathcal{E}}$, that are the *transition probabilities*;

or

- a probability space $(\Omega, \mathfrak{A}, P)$, a random variable X_0 with values in $\mathcal{E}$ and law $p^{(0)}$, and a *sequence of independent and identically distributed random variables* $(W_n)_{n \geq 0}$ with values in some discrete space $\mathcal{W}$, independent from X_0;
- with the *dynamics* at any time k, which is a map $f : \mathcal{A} \times \mathcal{W} \to \mathcal{E}$.

Consider also the following (stationary) parameters:

- the *instantaneous/running reward/payoff* (at any time k), which is a map $r : \mathcal{A} \to \mathbb{R}$, where, for all $x \in \mathcal{E}$ and $u \in \mathcal{C}(x)$, $r(x, u)$ denotes the reward of the action $u \in \mathcal{C}$ in state $x \in \mathcal{E}$;
- a (fixed) *discount factor* $\alpha \in [0, 1)$.

- for all strategies $\sigma = (\sigma_k)_{k\geq 0}$ in Σ or Σ^{R}, the *discounted total additive payoff* with infinite horizon:

$$J_\alpha^{(\sigma)} := J_\alpha(X;U) := \mathbb{E}\left[\sum_{k=0}^{\infty} \alpha^k r(X_k, U_k)\right] \ , \tag{4.1}$$

  where $(X,U) := (X_k, U_k)_{k\geq 0}$ is the process induced by σ as in Definition 3.2 or Definition 3.3.

- and the *discounted total additive payoff* with infinite horizon starting at x at time 0:

$$J_\alpha^{(\sigma)}(x) := J_{\alpha,x}(X;U) := \mathbb{E}\left[\sum_{k=0}^{\infty} \alpha^k r(X_k, U_k) \mid X_0 = x\right] \ ; \tag{4.2}$$

Given the above parameters (in particular a stationary MDP, a stationary reward and a discount factor), and a stationary feedback policy $\pi \in \Pi^{\mathrm{S}}$ (the set of stationary feedback policies), we associate the (stationary) *Markov transition matrix*: $M^{(\pi)}$ given by

$$M_{xy}^{(\pi)} := M_{xy}^{(\pi(x))}, \ \forall x, y \in \mathcal{E} \ .$$

We also associate the (stationary) *reward vector* $r^{(\pi)} \in \mathbb{R}^{\mathcal{E}}$ with

$$r_x^{(\pi)} := r(x, \pi(x)), \quad \text{for } x \in \mathcal{E} \ .$$

Fact 4.1. The associated stochastic process $(X_k, U_k)_{k\geq 0}$ (that is satisfying $U_k = \pi(X_k)$) is such that $(X_k)_{k\geq 0}$ is a stationary Markov chain with initial law $p^{(0)}$ and transition probability matrix $M^{(\pi)}$. Moreover, $(X_k, U_k)_{k\geq 0}$ is also a Markov chain taking its values in $\mathcal{E} \times \mathcal{C}$.

Using Kolmogorov equation for dicounted criteria (Theorem 2.25) and Fact 4.1, we deduce the following result.

Lemma 4.2. *When $\sigma = \pi$ is a stationary feedback policy, then the value function $v := J_\alpha^{(\pi)}$ satisfies the Kolmogorov equation:*

$$v = r^{(\pi)} + \alpha M^{(\pi)} v \ .$$

Definition 4.3. A *Markov decision problem with complete observation and infinite horizon discounted criteria* consists in the following optimization problem:

$$\max_{\sigma} J_\alpha^{(\sigma)}$$

where the optimization holds over either all relaxed strategies $\sigma \in \Sigma^{\mathrm{R}}$, or all pure strategies, or all Markov strategies, or all feedback policies, or all stationary feedback policies.

The optimum of above criteria is called the *value* of the problem.

An optimal solution σ is called an *optimal strategy*, and the corresponding process U_k or (X_k, U_k) an *optimal control process*.

Moreover, maximization can be replaced by minimization.

4.1.1 The stationary dynamic programming equation

Definition 4.4. For all $x \in \mathcal{E}$, let $v_\alpha(x)$ or simply $v(x)$ be the value of the Markov decision problem with an initial state x:

$$\max_\sigma J_\alpha^{(\sigma)}(x) \ .$$

The map $v_\alpha : \mathcal{E} \to \mathbb{R}, x \mapsto v_\alpha(x)$ is called the *value function* of the MDP.

Theorem 4.5 (Dynamic programming equation for Markov decision problems with discounted infinite horizon criteria)**.** *Assume that the map r is bounded from above. Let v be the value function of the Markov decision problem:*

$$v(x) := \max_\sigma J_\alpha^{(\sigma)}(x) \ ,$$

where the maximum is taken over all relaxed strategies (starting at time 0). Then, v is the unique solution of the following fixed point equation, called the stationary Bellman dynamic programming equation*:*

$$v(x) = \sup_{u \in \mathcal{C}(x)} \left(r(x,u) + \alpha \sum_{y \in \mathcal{E}} M_{xy}^{(u)} v(y) \right) \quad \forall x \in \mathcal{E} \ . \tag{4.3}$$

Moreover, the values v obtained by optimizing over the restricted sets of pure strategies, Markov strategies, feedback policies, or stationary feedback policies coincide.

Assume in addition that the maximum of (4.3) *is attained for an action $u \in \mathcal{C}(x)$ and* let us denote by $\pi(x)$ this action*, then the stationary feedback policy π (that is $(\pi_k)_{k\geq 0}$ with $\pi_k = \pi$) is an optimal strategy of the problem.*

We also have $w = p^{(0)}v$ for the value w of the Markov decision problem with infinite horizon discounted criteria.

The right hand side of the Bellman equation (4.3):

$$v(x) = \sup_{u \in \mathcal{C}(x)} \left(r(x,u) + \alpha \sum_{y \in \mathcal{E}} M_{xy}^{(u)} v(y) \right) \quad \forall x \in \mathcal{E}$$

can also be written as:

$$\sup_{u \in \mathcal{C}(x)} \Big(r(x,u) + \alpha \mathbb{E}\left[v(X_1) \mid X_0 = x, U_0 = u \right] \Big) \ .$$

Moreover, in the case of the definition of a MDP using a i.i.d. random sequence (W_n) and a dynamics f such that $X_{n+1} = f(X_n, U_n, W_n)$, the right hand side of the Bellman equation writes:

$$\sup_{u \in \mathcal{C}(x)} \Big(r(x,u) + \alpha \mathbb{E}\left[v(f(x,u,W_0)) \right] \Big) \ .$$

4.1.2 The Bellman operator

Definition 4.6. Let $\mathcal{B}_\alpha : \mathbb{R}^{\mathcal{E}} \to \mathbb{R}^{\mathcal{E}}$ be the map such that for all $v \in \mathbb{R}^{\mathcal{E}}$, and $x \in \mathcal{E}$, we have

$$[\mathcal{B}_\alpha(v)](x) = \sup_{u \in \mathcal{C}(x)} \left(r(x,u) + \alpha \sum_{y \in \mathcal{E}} M_{xy}^{(u)} v(y) \right) \ .$$

The map $\mathcal{B}_\alpha$ is called the *Bellman operator* of the discounted infinite horizon Markov decision problem.

The dynamic programming equation can be rewritten in functional form as the *fixed point equation* of the Bellman operator $\mathcal{B}_\alpha$:

$$v = \mathcal{B}_\alpha(v) \ .$$

Fact 4.7. The undiscounted Bellman operator $\mathcal{B}_1$ is monotone and additively homogenous.

Recall the proof given in the finite horizon case:
$\mathcal{B}_\alpha$ is the supremum of Kolmogorov operators

$$\mathcal{B}_\alpha^{(\pi)} : v \mapsto r^{(\pi)} + \alpha M^{(\pi)} v \ .$$

The operators $\mathcal{B}_1^{(\pi)}$ are monotone and additively homogeneous, so is $\mathcal{B}_1$. One can also do the proof by "hand" on:

$$\mathcal{B}_1(v)(x) = \sup_{u \in \mathcal{C}(x)} \Big(r(x,u) + \mathbb{E}\left[v(f(x,u,W_0)) \right] \Big) \ .$$

Expectation is monotone and it is additively homogenous because $\mathbb{E}[1] = 1$.

So is the previous expression as a function of v.

Corollary 4.8. *The discounted Bellman operator $\mathcal{B}_\alpha$ is Lipschitz continuous for the sup-norm with Lipschitz constant α, thus it is α-contracting when $\alpha < 1$.*

Corollary 4.9. *When $\mathcal{E}$ is finite and $\alpha < 1$, the operator $\mathcal{B}_\alpha$ admits a unique fixed point v^*. Moreover, for any initial point $v_0 \in \mathbb{R}^{\mathcal{E}}$, the sequence $v_{n+1} = \mathcal{B}_\alpha(v_n)$ converges towards v^*:*

$$\|v_n - v^*\|_\infty \leq \alpha^n \|v_0 - v^*\|_\infty \ .$$

Definition 4.10. The algorithm constructing the sequence $v_{n+1} = \mathcal{B}_\alpha(v_n)$ is called *value iterations*.

4.1.3 Proof of the Stationary Dynamic programming equation

1. v is solution of the Bellman equation. Assume that $\alpha < 1$. Let v^* be the unique solution of the Bellman equation $v = \mathcal{B}_\alpha(v)$, (by Fact 4.9).

Let $v^{(N)}$ be the value function of the finite horizon problem:

$$v^{(N)}(x) = \max_\sigma J_0^{(N,\sigma)}(x) \ ,$$

with

$$J_0^{(N,\sigma)}(x) := J_{0,x}^N(X;U) := \mathbb{E}\left[\left(\sum_{k=0}^{N-1} \alpha^k r(X_k, U_k) \right) + 0 \mid X_0 = x \right] \ .$$

From Theorem 3.13, $v^{(N)} = v_0^{(N)}$ with $v_k^{(N)}$ solution of the dynamic programming equation:

$$\begin{aligned} &v_N^{(N)}(x) = 0 \quad \forall x \in \mathcal{E} \ , \\ &v_k^{(N)}(x) = \sup\{\alpha^k r(x,u) + \sum_{y \in \mathcal{E}} M_{xy}^{(u)} v_{k+1}^{(N)}(y) \mid u \in \mathcal{C}(x)\} \quad \forall x \in \mathcal{E},\ k \leq N-1. \end{aligned}$$

This can be rewritten as

$$v_k^{(N)} = \alpha^k \mathcal{B}_\alpha(v_{k+1}^{(N)}/\alpha^{k+1}) \ .$$

Hence, $v^{(N)} = \mathcal{B}_\alpha^N(0) := \mathcal{B}_\alpha \circ \cdots \circ \mathcal{B}_\alpha(0)$ (where the composition is done N times). Therefore, $\lim_{N\to\infty} v^{(N)} = v^*$ where the limit is uniform in $\mathcal{E}$ (limit for the sup-norm of $\mathbb{R}^{\mathcal{E}}$).

Let C be a bound of r. Then, for all strategies σ, the sum $\sum_{k=0}^{\infty} \alpha^k r(X_k, U_k)$ exists a.s. since $|\alpha^k r(X_k, U_k)| \leq C\alpha^k$ (the series is absolutely convergent). Hence the expectation $J_\alpha^{(\sigma)}(x)$ also exists and is bounded. Therefore

$$v(x) := \sup_\sigma J_\alpha^{(\sigma)}(x) = \sup_\sigma \mathbb{E}\left[\sum_{k=0}^{\infty} \alpha^k r(X_k, U_k) \mid X_0 = x\right] \in \mathbb{R}$$

exists and satisfies

$$\|v - v^{(N)}\|_\infty \leq \sum_{k=N}^{\infty} \alpha^k C = \alpha^N \frac{C}{1-\alpha} \ ,$$

so $\lim_{N\to\infty} v^{(N)} = v$, which implies that $v = v^*$.

2. Optimality. Note that the bound $\|v - v^{(N)}\|_\infty \leq \alpha^N \frac{C}{1-\alpha}$ holds for the maximization over all nonstationary types of strategies: relaxed strategies, pure strategies, Markov strategies, or feedback policies. Since the value $v^{(N)}$ is the same if we maximize over all these types of strategies, the same property holds for v. To get the equality with stationary feedback strategies, and the optimality of the feedback policy given by Bellman equation, one proceed as for the proof of the finite horizon case.

If π is the stationary policy obtained as in the theorem, that is if $u = \pi(x)$ is optimal in (4.3), then v satisfies the Kolmogorov equation associated to π, and so (by Theorem 2.25), $v(x) = J_\alpha^{(\pi)}(x)$ which shows that π is optimal. If $u = \pi(x)$ is only ε-optimal for (4.3), that is

$$v(x) - \varepsilon \leq \left(r(x,u) + \alpha \sum_{y\in\mathcal{E}} M_{xy}^{(u)} v(y)\right)$$

Then, $v - \varepsilon\mathbf{1} \leq \mathcal{B}_\alpha^{(\pi)}(v)$. Iterating this inequality and using that $\mathcal{B}_1^{(\pi)}$ is monotone and additively homogenous, which, with $\mathcal{B}_\alpha^{(\pi)}(v) = \mathcal{B}_1^{(\pi)}(\alpha v)$, implies that $\mathcal{B}_\alpha^{(\pi)}$ is monotone and satisfies $\mathcal{B}_\alpha^{(\pi)}(\lambda\mathbf{1} + v) = \alpha\lambda\mathbf{1} + \mathcal{B}_\alpha^{(\pi)}(v)$, we get

$$v \leq \varepsilon + \mathcal{B}_\alpha^{(\pi)}(v) \leq \varepsilon(1+\alpha) + [\mathcal{B}_\alpha^{(\pi)}]^2(v) \leq \cdots \leq \varepsilon\frac{1-\alpha^{n+1}}{1-\alpha} + [\mathcal{B}_\alpha^{(\pi)})]^{n+1}(v) \ .$$

Using that $\mathcal{B}_\alpha^{(\pi)}$ is contracting, we get that the limit of $[\mathcal{B}_\alpha^{(\pi)})]^{n+1}(v)$ is equal to the solution of the stationary Kolmogorov equation, and is thus equal to $J_\alpha^{(\pi)}$. Passing to the limit when $n \to \infty$, in the previous inequality, we deduce:

$$v \leq \frac{\varepsilon}{1-\alpha} + J_\alpha^{(\pi)} \ .$$

This shows that π is then $\varepsilon/(1-\alpha)$-optimal for the Markov decision problem.

Since this holds for all $\varepsilon > 0$ (with a different π), we obtain that v is less or equal to the supremum of $J_\alpha^{(\pi)}$ over all stationary feedback policies. Since v is the supremum of $J_\alpha^{(\sigma)}$ over all strategies, and is thus greater or equal to the one over all stationary feedback policies, we deduce that both suprema are equal. □

4.2 Algorithms

4.2.1 Value iteration algorithm

Recall that the *value iteration* is the algorithm constructing

$$v_{n+1} = \mathcal{B}_\alpha(v_n) \ .$$

that is the fixed point iterations associated to the α-contracting operator $\mathcal{B}_\alpha$.

- It satisfies
$$\|v_n - v\|_\infty \leq \alpha^n \|v_0 - v\|_\infty \ .$$

- Denote $r_{\max}$ a bound on r (on both sides). Using the definition of v (as a maximum of the criterion (4.2), we found
$$\|v\|_\infty \leq \frac{r_{\max}}{1-\alpha} \ .$$
This can also be obtained from (4.3).

- To find a ε-solution, starting from 0, one need n iterations with $\alpha^n \frac{r_{\max}}{1-\alpha} \leq \varepsilon$, so the complexity is in
$$\mathcal{O}\Big(\frac{\log(\frac{(1-\alpha)\varepsilon}{r_{\max}})}{\log(\alpha)}\Big)nm \ ,$$
where $m = \mathrm{card}(\mathcal{A})$ and $n = \mathrm{card}(\mathcal{E})$ ($\mathcal{O}(nm)$ is the maximal complexity of the computation of $\mathcal{B}_\alpha(v)$ for some v).

- If $\alpha = 1 - \eta$ with $\eta = p/q$ a small rational, then the *length* (number of bit) of α is in the order of $\log(q)$, whereas $-1/\log(\alpha)$ is in the order of q so is *exponential* in the length of α, so in the length of the *input*.

- Hence value iteration is only *pseudo-polynomial*.

- If α is fixed, and we consider that the entries ε, $r(x,u)$ and $M_{xy}^{(u)}$ are rational numbers, we obtain that the complexity of value iteration is polynomial in the total length (number of bit) of the entries, so is a polynomial algorithm. Since the number of iterations depends on $r_{\max}$, value iteration algorithm is not strongly polynomial.

- In practice one uses rather a variant similar to Gauss-Seidel algorithm (wrt to Jacobi) for the solution of linear systems, in order to avoid useless storage. The resulting algorithm is called *Ford-Bellman algorithm*. It depends on some ordering on $\mathcal{E}$.

  It has a similar convergence rate and complexity.

- When $\alpha < 1$, or when $\alpha = 1$ and the MDP is not deterministic, *none of them converge in finite time*.

- This is already the case with no control: $\mathcal{C} = \{1\}$.

- For $\alpha < 1$, take $\mathcal{E} = \{1\}$, $r = r_{\max}$, and $v_0 = 0$, then $v_n \in \mathbb{R}$ satisfies

$$v_{n+1} = r_{\max} + \alpha v_n \implies v_n = r_{\max} \frac{1-\alpha^n}{1-\alpha}$$

  In this case, the number of iterations is equal to $\frac{\log(\frac{(1-\alpha)\varepsilon}{r_{\max}})}{\log(\alpha)}$. So the upper bound was tight.

- For $\alpha = 1$, take $\mathcal{E} = \{1, 2\}$, $M = \begin{bmatrix} 1/2 & 1/2 \\ 0 & 1 \end{bmatrix}$ and $r = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$. Then $v_n \in \mathbb{R}^2$ satisfies

$$v_{n+1}(1) = 1 + v_n(1)/2 + v_n(2)/2, \quad v_{n+1}(2) = v_n(2).$$

- If $v_0(2) = 0$, we get $v_n(1) = 2(1 - 2^{-n})$, and so the number of iterations is equal to $\frac{\log(\varepsilon/2)}{\log(1/2)}$.

4.2.2 Policy iteration algorithm

Recall that we want to solve

$$v(x) = \sup_{u \in \mathcal{C}(x)} \left(r(x,u) + \alpha \sum_{y \in \mathcal{E}} M_{xy}^{(u)} v(y) \right) \quad \forall x \in \mathcal{E} \ .$$

and find an optimal strategy of the Markov decision problem.

Assume that the action space $\mathcal{C}$ are finite, then for all $v \in \mathbb{R}^{\mathcal{E}}$, and $x \in \mathcal{E}$, the maximum in

$$\sup_{u \in \mathcal{C}(x)} \left(r(x,u) + \alpha \sum_{y \in \mathcal{E}} M_{xy}^{(u)} v(y) \right) \tag{4.4}$$

is attained by some action $\pi(x) \in \mathcal{C}(x)$.

Theorem 4.5 shows that computing π with respect to the solution of the Bellman equation yields a feedback policy which is optimal among all strategies.

The following algorithm has been introduced by Howard (1960) and is thus also called *Howard algorithm*.

Definition 4.11. The *policy iteration algorithm* applied to the Bellman equation $v = \mathcal{B}_\alpha(v)$ consists in the following successive steps $k \geq 0$, starting from a policy $\pi^0 \in \Pi$:

1. w^k is the unique solution of the Kolmogorov equation associated to the policy π^k:

$$v(x) = r(x, \pi^k(x)) + \alpha \sum_{y \in \mathcal{E}} M_{xy}^{(\pi^k(x))} v(y) \quad \forall x \in \mathcal{E} \ .$$

2. π^{k+1} is an optimal policy for w^k, that is an element π such that

$$\pi(x) \in \operatorname*{Argmax}_{u \in \mathcal{C}(x)} \left(r(x,u) + \alpha \sum_{y \in \mathcal{E}} M_{xy}^{(u)} w^k(y) \right) \quad \forall x \in \mathcal{E} \ .$$

The Bellman equation can be put in the form:

$$v = \mathcal{B}_\alpha(v) \ ,$$

where $\mathcal{B}_\alpha$ is the supremum of the Kolmogorov operators:

$$\mathcal{B}_\alpha^{(\pi)} : v \mapsto r^{(\pi)} + \alpha M^{(\pi)} v \ .$$

If π is optimal for v in (4.4), then

$$\mathcal{B}_\alpha(v) = \mathcal{B}_\alpha^{(\pi)}(v)$$

which means that

$$\mathcal{B}_\alpha(v) = \max_{\pi \in \Pi} \mathcal{B}_\alpha^{(\pi)}(v) \ .$$

Fact 4.12. The *policy iteration algorithm* applied to the Bellman equation $v = \mathcal{B}_\alpha(v)$ consists in the following successive steps, starting from a policy $\pi^0 \in \Pi$: for $k \geq 0$, do

1. w^k is the value of problem when the policy is freezed to π^k that is the solution of the equation $w = \mathcal{B}_\alpha^{(\pi^k)}(w)$.

2. π^{k+1} is an optimal policy for w^k, that is an element π of Π such that $\mathcal{B}_\alpha^{(\pi)}(w^k) = \mathcal{B}_\alpha(w^k)$.

Theorem 4.13. *Assume that the optimization problems in Bellman equations can be solved, that is, for all $v \in \mathbb{R}^{\mathcal{E}}$, there exists $\pi \in \Pi$ such that $\mathcal{B}_\alpha^{(\pi)}(v) = \mathcal{B}_\alpha(v)$. Denote by v the value function of the MDP with discounted criteria, that is the solution of $v = \mathcal{B}_\alpha(v)$.*

Then, for all $k \geq 0$, we have

$$w^k \leq w^{k+1} \leq \cdots \leq v \ ,$$

and

$$\lim_{k \to \infty} w^k = v \ .$$

Moreover,

$$w^k \leq \mathcal{B}_\alpha(w^k) \leq w^{k+1} \ ,$$

which means that the policy iteration algorithm converges faster than the value iteration algorithm, and we have

$$\|w^{k+1} - v\|_\infty \leq \alpha \|w^k - v\|_\infty \ .$$

The proofs of Policy iteration properties use the following properties that are of independent interest.

Proposition 4.14 (Sub or supersolutions)**.** *Let $\mathcal{B}$ be a monotone operator from $\mathbb{R}^{\mathcal{E}}$ to itself, which is contracting for the sup-norm. Let v be the unique fixed point of $\mathcal{B}$. Then*

$$w \leq \mathcal{B}(w) \implies w \leq v \tag{4.5}$$

$$w \geq \mathcal{B}(w) \implies w \geq v \ . \tag{4.6}$$

Proof. From $w \leq \mathcal{B}(w)$ and monotonicity of $\mathcal{B}$, we get, for all $n \geq 1$, $w \leq \cdots \leq \mathcal{B}^n(w)$.

Since $\mathcal{B}$ is contracting, $\mathcal{B}^n(w)$ converges towards v, which implies $w \leq v$. □

Definition 4.15. A solution of $w \leq \mathcal{B}(w)$ is called a *subsolution* of Bellman equation.

A solution of $w \geq \mathcal{B}(w)$ is called a *supersolution* of Bellman equation.

Proof of Theorem 4.13. We have

$$w^k = \mathcal{B}_\alpha^{(\pi^k)}(w^k) \tag{4.7}$$

$$\mathcal{B}_\alpha(w^k) = \mathcal{B}_\alpha^{(\pi^{k+1})}(w^k) \tag{4.8}$$

$$w^{k+1} = \mathcal{B}_\alpha^{(\pi^{k+1})}(w^{k+1}) \ . \tag{4.9}$$

Using the first and second equation together with the definition of $\mathcal{B}_\alpha$, we get

$$w^k = \mathcal{B}_\alpha^{(\pi^k)}(w^k) \leq \mathcal{B}_\alpha(w^k) = \mathcal{B}_\alpha^{(\pi^{k+1})}(w^k) \ .$$

Hence w^k is a subsolution of the Kolmogorov equation $w = \mathcal{B}_\alpha^{(\pi^{k+1})}(w)$. From (4.5) applied to the operator $\mathcal{B}_\alpha^{(\pi^{k+1})}$, and (4.9), we deduce that $w^k \leq w^{k+1}$.

Since we also have $w^k \leq \mathcal{B}_\alpha(w^k)$, w^k is a subsolution of the Bellman equation $w = \mathcal{B}_\alpha(w)$. Applying (4.5) to $\mathcal{B}_\alpha$, we deduce that $w^k \leq v$, for all $k \geq 0$.

From the above inequalities, we also get that $w^k \leq \mathcal{B}_\alpha(w^k) = \mathcal{B}_\alpha^{(\pi^{k+1})}(w^k)$, and so $\mathcal{B}_\alpha^{(\pi^{k+1})}(w^k) \leq (\mathcal{B}_\alpha^{(\pi^{k+1})})^2(w^k) \leq \cdots \leq w^{k+1}$. This shows that $w^k \leq \mathcal{B}_\alpha(w^k) \leq w^{k+1} \leq v$, which is the second assertion. With this we get that

$$0 \leq v - w^{k+1} \leq v - \mathcal{B}_\alpha(w^k)$$

hence

$$\|v - w^{k+1}\|_\infty \leq \|v - \mathcal{B}_\alpha(w^k)\|_\infty = \|\mathcal{B}_\alpha(v) - \mathcal{B}_\alpha(w^k)\|_\infty \leq \alpha \|v - w^k\|_\infty \ .$$

This shows that the sequence w^k converges towards v and that the convergence is faster than the one of value iterations. □

Theorem 4.16. *Assume that the action spaces $\mathcal{C}(x)$ are all finite. Then, the policy iteration algorithm converges in a finite number of steps.*

Proof. Since the sets $\mathcal{C}(x)$ are finite, the set of feedback policies, Π is finite. Therefore, there exists $k < \ell$ such that $\pi^k = \pi^\ell$. From the uniqueness of the solution w to the equation $\mathcal{B}_\alpha^{(\pi^k)}(w) = w$, we get that $w^k = w^\ell$. Since the sequence w^k is nondecreasing, $w^k \leq w^{k+1} \leq \cdots \leq w^\ell$, and satisfies $w^k = w^\ell$, we get the equality $w^k = w^{k+1}$. Then, using the definition of w^{k+1} and π^{k+1}, and $w^k = w^{k+1}$, we deduce $w^k = w^{k+1} = \mathcal{B}_\alpha^{(\pi^{k+1})}(w^{k+1}) = \mathcal{B}_\alpha^{(\pi^{k+1})}(w^k) = \mathcal{B}_\alpha(w^k)$. Hence, w^k is a fixed point solution of $\mathcal{B}_\alpha$ and since $\mathcal{B}_\alpha$ is contracting, the fixed point is unique, so $w^k = v$. Moreover, since π^{k+1} satisfies $\mathcal{B}_\alpha(v) = \mathcal{B}_\alpha^{(\pi)}(v)$, then it is optimal, and $w^\ell = v$ for all $\ell \geq 1$. □

4.2.3 Additional properties of Policy iterations for discounted problems

In Theorem 4.18 below, we prove that the policy iterations coincide with Newton algorithm applied to the system of equations $v - \mathcal{B}_\alpha(v) = 0$.

Let us first remark that if the map $\mathcal{B}_\alpha$ is regular, then the latter Newton algorithm consists in the iterations $(w^k)_{k\geq 0}$ in which w^{k+1} is the solution of the tangent equation in point w^k, that is

$$D_{w^k}(I - \mathcal{B}_\alpha)(w^{k+1} - w^k) + (I - \mathcal{B}_\alpha)(w^k) = 0$$

which can be rewritten as:

$$w^{k+1} = \mathcal{B}_\alpha(w^k) + D_{w^k}(\mathcal{B}_\alpha)(w^{k+1} - w^k) \ .$$

Moreover, one can weaken Newton algorithm by replacing the above differential by a subdifferential of $\mathcal{B}_\alpha$. Therefore, using the following result, one can see that policy iteration algorithm belongs to the class of generalized Newton algorithms, since then the above equation is equivalent to

$$w^{k+1} = r^{\pi^{w^k}} + \alpha M^{(\pi^{w^k})} w^{k+1} = \mathcal{B}_\alpha^{\pi^{w^k}}(w^{k+1}) \ .$$

Lemma 4.17. *For all $x \in \mathcal{E}$, the map $\psi : v \in \mathbb{R}^{\mathcal{E}} \mapsto [\mathcal{B}_\alpha(v)](x) \in \mathbb{R}$ is convex, and for all $v \in \mathbb{R}^{\mathcal{E}}$, and $u \in \mathrm{Argmax}_{u \in \mathcal{C}(x)}\{r(x,u) + \alpha M_x^{(u)} v\}$, the row vector $\alpha M_x^{(u)}$ is in the subdifferential of ψ at point v.*

Proof. For all $x \in \mathcal{E}$, the map $\psi : v \in \mathbb{R}^{\mathcal{E}} \mapsto [\mathcal{B}_\alpha(v)](x) \in \mathbb{R}$ is convex as a supremum of affine maps: $\psi(x) = [\mathcal{B}_\alpha(v)](x) = \sup_{u \in \mathcal{C}(x)} r(x,u) + \alpha M_x^{(u)} v$. If $u \in \mathrm{Argmax}_{u \in \mathcal{C}(x)}\{r(x,u) + \alpha M_x^{(u)} v\}$, we get that

$$\psi(w) - \psi(v) = [\mathcal{B}_\alpha(w)](x) - [\mathcal{B}_\alpha(v)](x) \geq (r(x,u) + \alpha M_x^{(u)} w) - (r(x,u) + \alpha M_x^{(u)} v) = \alpha M_x^{(u)}(w - v) \ .$$

This shows that the row vector $\alpha M_x^{(u)}$ is in the subdifferential of ψ at point v. □

Theorem 4.18. *Assume that $\mathcal{E}$ is finite, that the sets $\mathcal{C}(x)$ are compact spaces, that there exists a continuous map $v \in \mathbb{R}^{\mathcal{E}} \mapsto \pi^v \in \Pi$ such that, for all $v \in \mathbb{R}^{\mathcal{E}}$, π^v is the unique element of Π such that $\mathcal{B}_\alpha^{(\pi^v)}(v) = \mathcal{B}_\alpha(v)$, and that the map $u \in \mathcal{C}(x) \mapsto M_{xy}^{(u)}$ is continuous for all $x, y \in \mathcal{E}$. Then, $\mathcal{B}_\alpha$ is of class $\mathcal{C}^1$, the policy iteration algorithm coincides with Newton algorithm associated to the fixed point equation $v = \mathcal{B}_\alpha(v)$, and its convergence is superlinear:*

$$\lim_{k \to \infty} \frac{\|w^{k+1} - v\|_\infty}{\|w^k - v\|_\infty} = 0 \ .$$

Proof. Assume that a continuous map $v \in \mathbb{R}^{\mathcal{E}} \mapsto \pi^v \in \Pi$ exists with the property that π^v is the unique element of Π such that $\mathcal{B}_\alpha^{(\pi^v)}(v) = \mathcal{B}_\alpha(v)$. This means that, for all $v \in \mathbb{R}^{\mathcal{E}}$, $x \in \mathcal{E}$, $u = \pi^v(x)$ is the unique element of $\mathrm{Argmax}_{u \in \mathcal{C}(x)}\{r(x,u) + \alpha M_x^{(u)} v\}$. Then, by Lemma 4.17, $\alpha M_x^{(u)}$ is in the subdifferential of $\psi : v \in \mathbb{R}^{\mathcal{E}} \mapsto [\mathcal{B}_\alpha(v)](x) \in \mathbb{R}$ at point v. Similarly, for all $w \in \mathbb{R}^{\mathcal{E}}$, $\alpha M_x^{(\pi^w(x))}$ is in the subdifferential of ψ at w. Let p be any element of the subdifferential of ψ at v, which means that for all $w \in \mathbb{R}^{\mathcal{E}}$, we have

$$\psi(w) - \psi(v) \geq p(w - v) \ .$$

Then, we also have

$$\psi(w) - \psi(v) \leq \alpha M_x^{(\pi^w(x))}(w - v) \ ,$$

since $\alpha M_x^{(\pi^w(x))}$ is in the subdifferential of ψ at w. Therefore

$$p(w - v) \leq \alpha M_x^{(\pi^w(x))}(w - v) \ .$$

Consider the sequence $w_n = v + \frac{1}{n} z$, we deduce

$$pz \leq \alpha M_x^{(\pi^{w_n}(x))} z \ .$$

Using that w_n converges towards v, and that the maps $w \mapsto \pi^w(x) \in \mathcal{C}(x)$ and $u \in \mathcal{C}(x) \mapsto M_{xy}^{(u)}$ are continuous, we deduce that $\alpha M_x^{(\pi^{w_n}(x))}$ converges towards $\alpha M_x^{(\pi^v(x))}$. This implies that $pz \leq \alpha M_x^{(\pi^v(x))} z$. Since this holds for all $z \in \mathbb{R}^{\mathcal{E}}$, we deduce that $p = \alpha M_x^{(\pi^v(x))}$. We have shown that the subdifferential of ψ is reduced to a singleton. Then, the map is differentiable [6] and $\alpha M_x^{(\pi^v(x))}$ is the differential of $v \mapsto [\mathcal{B}_\alpha(v)](x)$. Applying this property for all $x \in \mathcal{E}$, we get that $\mathcal{B}_\alpha$ si differentiable at v with differential equal to $\alpha M^{(\pi^v)}$. Since this differential is continuous with respect to v, this shows also that $\mathcal{B}_\alpha$ is of class $\mathcal{C}^1$. Then, the above comments show that policy iteration algorithm coincides with Newton algorithm.

Now let v be the value function. We have

$$(I - \alpha M^{(\pi^{w^k})})(w^{k+1} - v) = r^{\pi^{w^k}} + \alpha M^{(\pi^{w^k})} v - v \tag{4.10}$$

$$= \mathcal{B}_\alpha(w^k) - \mathcal{B}_\alpha(v) - \alpha M^{(\pi^{w^k})}(w^k - v) \ . \tag{4.11}$$

We can show

$$\|w^{k+1} - v\|_\infty \leq \frac{1}{1-\alpha} \|\mathcal{B}_\alpha(w^k) - \mathcal{B}_\alpha(v) - \alpha M^{(\pi^{w^k})}(w^k - v)\|_\infty \ .$$

Since we already proved that $\mathcal{B}_\alpha$ is of class $\mathcal{C}^1$ and that its differential at v is equal to $\alpha M^{(\pi^v)}$, we get that the right hand side of the above inequality is in $o(\|w^k - v\|_\infty)$, which gives the superlinear convergenve of Policy Iterations.

Another proof is using directly the above subdifferential properties to show from (4.10):

$$\alpha(M^{(\pi^v)} - M^{(\pi^{w^k})})(w^k - v) \leq (I - \alpha M^{(\pi^{w^k})})(w^{k+1} - v) \leq 0 \ .$$

Then, we obtain

$$\|w^{k+1} - v\|_\infty \leq \frac{\alpha}{1-\alpha} \|(M^{(\pi^{w^k})} - M^{(\pi^v)})(w^k - v)\|_\infty \ ,$$

and since the map $v \mapsto M^{(\pi^v)}$ is continuous, the right hand side of the above inequality is in $o(\|w^k - v\|_\infty)$, which gives the superlinear convergenve of Policy Iterations. □

Note that contrarilly to the general situation of the Newton algorithm, the policy iterations always converge towards the solution of the fixed point equation. This comes from the convexity of the maps $v \mapsto [\mathcal{B}(v)](x)$, which implies the monotonicity of the sequence of value functions w^k.

4.3 Optimal stopping time problems with infinite horizon

Consider

- a (fixed) stationary Markov chain $(X_n)_{n \geq 0}$ over a probability space $(\Omega, \mathfrak{A}, P)$ with values in a finite state space $\mathcal{E}$ and transition matrix M and initial probability law $p^{(0)}$.
- an *instantaneous/running reward/payoff* (at any time k), which is a map $r : \mathcal{A} \to \mathbb{R}$;
- a (fixed) *discount factor* $\alpha \in [0, 1)$.

- for all stopping times τ with respect to the Markov chain $(X_n)_{n\geq 0}$, the *discounted infinite horizon payoff with stopping time* τ:

$$J_\alpha^{(\tau)} := \mathbb{E}\left[\left(\sum_{\ell=0}^{\tau-1} \alpha^\ell r(X_\ell)\right) + \alpha^\tau \varphi(X_\tau)\right] \ ; \tag{4.12}$$

- and the *discounted infinite horizon payoff with stopping time* τ, starting in x at time 0:

$$J_\alpha^{(\tau)} := \mathbb{E}\left[\left(\sum_{\ell=0}^{\tau-1} \alpha^\ell r(X_\ell)\right) + \alpha^\tau \varphi(X_\tau) \mid X_0 = x\right] \ . \tag{4.13}$$

Definition 4.19. An *Optimal stopping time problem with complete observation and infinite horizon discounted criteria* consists in the following optimization problem:

$$\max_\tau J_\alpha^{(\tau)}$$

where the optimization holds over all stopping times τ with respect to the Markov chain $(X_n)_{n\geq 0}$.

The optimum of above criteria is called the *value* of the problem.

An optimal solution τ is called an *optimal stopping time*.

Definition 4.20. For all $x \in \mathcal{E}$, let $v_\alpha(x)$ or simply $v(x)$ be the value of the optimal stopping time problem with an initial state x:

$$\max_\tau J_\alpha^{(\tau)}(x) \ .$$

The map $v_\alpha : \mathcal{E} \to \mathbb{R}, x \mapsto v_\alpha(x)$ is called the *value function* of the stopping time problem.

Theorem 4.21 (Dynamic programming equation for optimal stopping time problems with discounted infinite horizon criteria)**.** *Assume that the map r is bounded from above. Let v be the value function of the optimal stopping time problem:*

$$v(x) := \max_\tau J_\alpha^{(\tau)}(x) \ ,$$

where the maximum is taken over all stopping times τ with respect to the Markov chain $(X_n)_{n\geq 0}$. Then, v is the unique solution of the following fixed point equation, called stationary Bellman equation *or* variational inequality*:*

$$v(x) = \max\left(r(x) + \alpha \sum_{y\in\mathcal{E}} M_{xy} v(y), \varphi(x)\right) \quad \forall x \in \mathcal{E} \ . \tag{4.14}$$

Let B be the set of states in which the maximum in (4.14) *is attained in the first term, that is*

$$B := \{x \in \mathcal{E} \mid r(x) + \alpha \sum_{y\in\mathcal{E}} M_{xy} v(y) \geq \varphi(x)\} \ .$$

Then an optimal stopping time τ is obtained by choosing for τ the exit time τ_B from B of the Markov chain.

Proof. Consider the MDP in which

- the state space is $\mathcal{E}' = \mathcal{E} \cup \{c\}$ where $c \notin \mathcal{E}$;
- the control space $\mathcal{C} = \{0, 1\}$ (0 for stop, and 1 for not stop);
- the control space $\mathcal{C}(x)$ is such that $\mathcal{C}(x) = \mathcal{C}$ if $x \in \mathcal{E}$ and $\mathcal{C}(x) = \{0\}$ if $x = c$.
- the states of the MDP, Y_n, depend on the states of the Markov chain X_n and on the actions as follows:

$$Y_{n+1} = g(X_{n+1}, U_n)$$

where $g(x, u) = x$ if $u = 1$ and $g(x, u) = c$ otherwise.

Then, for all $x_i \in \mathcal{E}'$, $u_i \in \mathcal{C}_i(x_i)$, $i \geq 0$, we have

$$\begin{aligned} &P(Y_{k+1} = x_{k+1} \mid Y_k = x_k, U_k = u_k, Y_{k-1} = x_{k-1}, \ldots, Y_0 = x_0, U_0 = u_0) \\ &\quad = 1 \text{ if } c = x_{k+1} \text{ and } u_k = 0 \\ &\quad = 0 \text{ if } x_{k+1} \in \mathcal{E} \text{ and } u_k = 0 \\ &\quad = M_{x_k, x_{k+1}} \text{ if } x_k, x_{k+1} \in \mathcal{E} \text{ and } u_k = 1 \end{aligned}$$

This implies that

$$\begin{aligned} &P(Y_{k+1} = x_{k+1} \mid Y_k = x_k, U_k = u_k, Y_{k-1} = x_{k-1}, \ldots, Y_0 = x_0, U_0 = u_0) \\ &\quad = P(Y_{k+1} = x_{k+1} \mid Y_k =_k, U_k = u_k) \ . \end{aligned}$$

which is the Markov property. The transition vectors of the MDP are: $M^{(1)}_{xy} = M_{xy}$ for all $x, y \in \mathcal{E}$, $M^{(1)}_{xy} = 0$ for all $y = c$, and $M^{(0)}_{xy} = 1$ for $y = c$ and 0 for $y \in \mathcal{E}$.

Let us take the rewards: $r'(x, 1) = r(x)$, $r'(x, 0) = \phi(x)$ for $x \in \mathcal{E}$ and $r'(c, 0) = 0$.

Then, the value of the discounted infinite horizon problem coincides with the value of the stopping time problem. Indeed, take $\tau = \inf\{t \geq 0 \mid U_t = 0\}$, and conversely take $U_n = 1$ for all $n < \tau$ and $U_n = 0$ for $n \geq \tau$. Then, τ is a stopping time if and only if U_n is given by a strategy.

The Bellman equation of the Markov decision problem is then:

$$\begin{aligned} &v(x) = \max\Big(r(x) + \alpha \sum_{y \in \mathcal{E}} M_{xy} v(y), \varphi(x) + v(c)\Big) \quad \forall x \in \mathcal{E} \ , \\ &v(c) = 0 \ . \end{aligned}$$

Here the action $u = 1$ corresponds to the left term in the above maximum, and $u = 0$ corresponds to the right term. Therefore if $\pi : \mathcal{E} \to \{0, 1\}$ is an optimal policy given by the Bellman equation, we recover again the optimal stopping time by taking:

$$\tau = \inf\{t \geq 0 \mid \pi(X_t) = 0\}$$

or by taking $\tau = \tau_B$ where $B = \{x \in \mathcal{E} \mid \pi(x) = 1\}$. □

4.4 Problems with variably discounted infinite horizon payoff

Assume given a stationary Markov decision process, and the following stationary parameters:

- the *instantaneous/running reward/payoff* (at any time k), which is a map $r : \mathcal{A} \to \mathbb{R}$;
- a variable *discount factor* (at any time k), which is a map $\alpha : \mathcal{A} \to \mathbb{R}_+$;
- for all strategies $\sigma = (\sigma_k)_{k\geq 0}$ in Σ or Σ^{R}, the *variably discounted total additive payoff* with infinite horizon:

$$J^{(\sigma)} := J(X;U) := \mathbb{E}\left[\sum_{\ell=0}^{\infty}\left(\prod_{m=0}^{\ell-1}\alpha(X_m,U_m)\right) r(X_\ell,U_\ell)\right] \ , \tag{4.15}$$

  where $(X,U) := (X_k,U_k)_{k\geq 0}$ is the process induced by σ as in Definition 3.2 or Definition 3.3.
- and the *variably discounted total additive payoff* with infinite horizon, starting x at time 0:

$$J^{(\sigma)}(x) := J_x(X;U) := \mathbb{E}\left[\sum_{\ell=0}^{\infty}\left(\prod_{m=0}^{\ell-1}\alpha(X_m,U_m)\right) r(X_\ell,U_\ell) \mid X_k = x\right] \ , \tag{4.16}$$

Theorem 4.22 (Dynamic programming equation for Markov decision problems with variably discounted infinite horizon criteria)**.** *Assume that the map r is bounded from above and that $\alpha(x,u) \leq \bar{\alpha}$ for all $(x,u) \in \mathcal{A}$, for some constant $\bar{\alpha} < 1$. Let v be the value function of the Markov decision problem associated to the above parameters:*

$$v(x) := \max_{\sigma} J^{(\sigma)}(x) \ ,$$

where the maximum is taken over all relaxed strategies (starting at time 0). Then, v is the unique solution of the following fixed point equation, called the stationary Bellman dynamic programming equation*:*

$$v(x) = \sup_{u\in\mathcal{C}(x)}\left(r(x,u) + \alpha(x,u)\sum_{y\in\mathcal{E}} M_{xy}^{(k,u)} v(y)\right) \quad \forall x \in \mathcal{E} \ . \tag{4.17}$$

Moreover, the values v obtained by optimizing over the restricted sets of pure strategies, Markov strategies, or feedback policies, or stationary feedback policies coincide.

Assume in addition that the maximum of (4.17) *is attained for an action $u \in \mathcal{C}(x)$ and* let us denote by $\pi(x)$ this action*, then the stationary feedback policy π (that is $(\pi_k)_{k\geq 0}$ with $\pi_k = \pi$) is an optimal strategy of the problem. problem.*

The arguments of the proof are the same as when α is constant. Let us consider the Bellman operator of the variably discounted problem which is the map $\mathcal{B} : \mathbb{R}^{\mathcal{E}} \to \mathbb{R}^{\mathcal{E}}$ such that, for all $v \in \mathbb{R}^{\mathcal{E}}$, and $x \in \mathcal{E}$, we have

$$[\mathcal{B}(v)](x) = \sup_{u\in\mathcal{C}(x)}\left(r(x,u) + \alpha(x,u)\sum_{y\in\mathcal{E}} M_{xy}^{(u)} v(y)\right) \ .$$

The map $\mathcal{B}$ satisfies the following properties

Lemma 4.23. *Under the assumptions of Theorem 4.22, $\mathcal{B}$ is monotone and $\bar{\alpha}$-additively subhomogenous (Definition 3.22), meaning that for all $v \in \mathbb{R}^{\mathcal{E}}$ and $\lambda \geq 0$, we have*

$$\mathcal{B}(v + \lambda \mathbf{1}) \leq \mathcal{B}(v) + \bar{\alpha}\lambda \mathbf{1} \ .$$

Therefore, it is Lipschitz continuous for the sup-norm with Lipschitz constant $\bar{\alpha} < 1$, thus it is $\bar{\alpha}$-contracting.

Proof. The first assertion can be proved elementarily. Let us prove the second one using the first one. Let $v, v' \in \mathbb{R}^{\mathcal{E}}$. Denote $\lambda = \|v - v'\|_\infty$. We have $v(x) \leq \lambda + v'(x)$ for all $x \in \mathcal{E}$, that is $v \leq v' + \lambda \mathbf{1}$. Since $\mathcal{B}$ is monotone, we get $\mathcal{B}(v) \leq \mathcal{B}(v' + \lambda \mathbf{1})$. Since $\mathcal{B}$ is additively subhomogenous with constant $\bar{\alpha}$, we deduce $\mathcal{B}(v' + \lambda \mathbf{1}) \leq \mathcal{B}(v') + \bar{\alpha}\lambda$. Then, $\mathcal{B}(v) \leq \mathcal{B}(v') + \bar{\alpha}\lambda \mathbf{1}$, hence $\max_{x \in \mathcal{E}} [\mathcal{B}(v)](x) - [\mathcal{B}(v')](x) \leq \bar{\alpha}\lambda$. Exchanging v and v' we get the other inequality: $\max_{x \in \mathcal{E}} [\mathcal{B}(v')](x) - [\mathcal{B}(v)](x) \leq \bar{\alpha}\lambda$. Hence $\|\mathcal{B}(v') - \mathcal{B}(v)\|_\infty = \max_{x \in \mathcal{E}} \max([\mathcal{B}(v')](x) - [\mathcal{B}(v)](x), [\mathcal{B}(v)](x) - [\mathcal{B}(v')](x)) \leq \bar{\alpha}\lambda = \bar{\alpha}\|v' - v\|_\infty$. This shows the Lipschitz continuity. □

Proof of Theorem 4.22. We use the same technique as for the constant discount factor case: first use of Bellman equation for finite horizon problems with mixed criterias, then take the limit using contraction, then use of the stationary Kolmogorov equation for infinite horizon criteria with variable discount factor. □

Exercise 4.4.1. Show that one can reduce the above problem to an infinite horizon problem with constant discount factor equal to $\bar{\alpha}$, by adding a cemetery point to the state space $\mathcal{E}$.

Corollary 4.24. *Under the assumptions of Theorem 4.22, the sequence $v^{(T)}$ of value functions of finite horizon problems with mixed criteria (given in Remark 3.21) converges when T goes to infinity to the unique solution of the stationary Bellman dynamic programming equation* (4.17).

4.5 Problems with exit time in infinite horizon

Assume given a stationary Markov decision process, and the following stationary parameters:

- a strict subset B of $\mathcal{E}$;
- a *final reward*, which is a map $\varphi : \mathcal{E} \to \mathbb{R}$;
- a (fixed) *discount factor* $\alpha \in [0, 1)$;
- the *instantaneous/running reward/payoff* (at any time k), which is a map $r : \mathcal{A} \to \mathbb{R}$;
- for all strategies $\sigma = (\sigma_k)_{k \geq 0}$ in Σ or Σ^{R}, the *payoff with exit time* in infinite horizon:

$$J^{(B,\sigma)} := J^B(X; U) := \mathbb{E}\left[\sum_{\ell=0}^{\tau_B - 1} \alpha^\ell r(X_\ell, U_\ell) + \alpha^{\tau_B} \varphi(X_{\tau_B}) \mathbf{1}_{\tau_B < +\infty}\right] \ , \tag{4.18}$$

  where $(X, U) := (X_k, U_k)_{k \geq 0}$ is the process induced by σ as in Definition 3.2 or Definition 3.3, and τ_B is the exit time of the process $(X_n)_{n \geq 0}$ from B.

- and the *payoff with exit time* and infinite horizon, starting in x at time 0:

$$J^{(B,\sigma)}(x) := J_x^B(X;U) := \mathbb{E}\left[\sum_{\ell=0}^{\tau_B-1} \alpha^\ell r(X_\ell, U_\ell) + \alpha^{\tau_B}\varphi(X_{\tau_B})\mathbf{1}_{\tau_B<+\infty} \mid X_0 = x\right] \quad . \tag{4.19}$$

Theorem 4.25 (Dynamic programming equation for Markov decision problems with exit time in discounted infinite horizon)**.** *Assume that the maps φ, r are bounded from above. Let v be the value function of the Markov decision problem:*

$$v(x) := \max_\sigma J^{(B,\sigma)}(x) \ ,$$

where the maximum is taken over all relaxed strategies (starting at time 0). Then, v is the unique solution of the following fixed point equation, called the stationary Bellman dynamic programming equation*:*

$$v(x) = \sup_{u\in\mathcal{C}(x)} \left(r(x,u) + \alpha \sum_{y\in\mathcal{E}} M_{xy}^{(k,u)} v(y) \right) \quad \forall x \in B \ . \tag{4.20a}$$

with boundary condition

$$v(x) = \varphi(x), \quad \forall x \notin B \ . \tag{4.20b}$$

Moreover, the values v obtained by optimizing over the restricted sets of pure strategies, Markov strategies, or feedback policies, coincide.

Assume in addition that the maximum of (4.20) *is attained for an action $u \in \mathcal{C}(x)$, for $x \in B$, and* let us denote by $\pi(x)$ this action *when $x \in B$, and choose any action $\pi(x)$ for $x \notin B$, then the stationary feedback policy π (that is $(\pi_k)_{k\geq 0}$ with $\pi_k = \pi$) is an optimal strategy of the problem.*

As for finite horizon problems, Theorem 4.25 can be deduced easily from Theorem 4.22 using the following property which is the same as Fact 2.34 and Fact 3.26.

Fact 4.26. The functional $J^{(B,\sigma)}$ of (4.19) can rewritten as the infinite horizon mixed functional:

$$J^{(B,\sigma)}(x) := J_x^B(X;U) := \mathbb{E}\left[\sum_{\ell=0}^{\infty} \left(\prod_{m=0}^{\ell-1} \alpha(X_m, U_m)\right) r'(X_\ell, U_\ell)\right] \ ,$$

for the same Markov Decision process, with the instantaneous rewards r' and variable discount factors α given by:

$$\begin{aligned}
r'(x,u) &= r(x,u), \quad \text{for } x \in B,\ u \in \mathcal{C}(x) \\
r'(x,u) &= \varphi(x), \quad \text{for } x \notin B,\ u \in \mathcal{C}(x) \\
\alpha(x,u) &= \alpha, \quad \text{for } x \in B,\ u \in \mathcal{C}(x) \\
\alpha(x,u) &= 0, \quad \text{for } x \notin B,\ u \in \mathcal{C}(x) \ .
\end{aligned}$$

4.6 Problem: Divorce of Birds

This problem is taken from the (ENSTA+M2) exam of 2016/2017.

We consider the modelization of the decision of divorce of birds as a MDP. We assume that at each breeding (reproduction) season, the bird female has a mate, and that at the end of the season, she is taking the decision on whether to divorce her mate. Then, winter arrives and the female and the male may die. If the female has no mate after winter (she divorced, or the male died, or it is the first breeding season of the female), then she is choosing a mate among a "pool" of available males. The decision of the female is based on the qualities of the male and female, and also on the information on whether it is the first breeding season of the female, or the female divorced, or the male died during winter.

We consider the following notations, parameters and assumptions:

- We consider one female during her life, and denote by $Y_k \in \mathcal{Y} \subset \mathbb{R}$ her quality at the begining of the breeding season of year k (one can start to number years after the female is able to breed). We shall assume that $\mathcal{Y} = [\bar{y}] := \{0, \dots, \bar{y}\}$, where $y = 0$ means that the female is dead. We denote by $\mathcal{Y}^* = \mathcal{Y} \setminus \{0\}$.
- We denote by $X_k \in \mathcal{X} \subset \mathbb{R}$ the quality of the male choosen by the female at the begining of the breeding season of year k. Again, one may assume that $\mathcal{X} = [\bar{x}]$.
- We denote by $Z_k \in \mathcal{Z} = \{0, 1, M\}$ the information on whether it is the first breeding season of the female or the female divorced (in which case $Z_k = 0$), or the male died during winter (in which case $Z_k = M$), or the female mated with the same male the previous year (in which case $Z_k = 1$).
- We denote by $U_k \in \mathcal{C} = \{0, 1\}$ the decision of the female to divorce: $U_k = 1$ if she decides to divorce and 0 otherwise.
- $r(x, y, z)$ is the reproductive success, that is the expected number of children, during one season, when the qualities of the male and female are x and y and the information on whether it is the first breeding season of the female or the female divorced or the male died during winter is z.
- We assume that the pools of males in which the female is choosing a partner each year when needed are independent and that $f(x)$ is the probability of finding a male with quality x in a pool.
- s_f and s_m are respectively the survival probabilities of a female and a male after winter. They are independent of age, constant and < 1.
- We assume that the quality of males and females do not vary with time until their death.

Q 6.1. Show that the sequence Y_k is a Markov chain and compute its transition probability.

Q 6.2. Show that the sequences X_k, Y_k, Z_k, U_k define a MDP, precise what is the state and what is the control, and compute the transition probabilities.

Q 6.3. We assume that the aim of the female is to maximize her reproductive success, that is the expected total number of her children, during all her life. Write this as an infinite horizon criterion for the MDP.

Q 6.4. Let $v(x, y, z)$ be the value of the previous problem when the initial qualities of male and female are $x \in \mathcal{X}$ and $y \in \mathcal{Y}^*$ and the information on what happened before and during previous winter is $z \in \mathcal{Z}$. What is the equation satisfied by the function v? How to find an optimal policy?

Q 6.5. Show that the equation of v is of the form $v(x, y, z) = [F(v)](x, y, z)$ where F is an operator on the set of functions from $\mathcal{X} \times \mathcal{Y}^* \times \mathcal{Z}$ to $\mathbb{R}$ of the form:

$$[F(v)](x, y, z) = r(x, y, z) + s_f \max\Big([M^{(1)}v](y), [M^{(0)}v](x, y)\Big)$$

where v is identified to a vector, for $u = 0, 1$, $M^{(u)}$ is a Markov matrix, and $[M^{(0)}v]_{x,y,z}$ does not depend on z and $[M^{(1)}v]_{x,y,z}$ does not depend on x and z.

Q 6.6. Deduce that the operator F is contracting for the sup-norm $\|v\| = \max_{x,y,z} |v(x, y, z)|$.

Q 6.7. Show that the fixed point v of F is unique.

Q 6.8. Show that if r is nondecreasing with respect to x, then so does v.

Q 6.9. Show in that case that there exists $x^*(y)$ such that the optimal policy of the female y is to divorce from the male x if and only if $x < x^*(y)$.

Some references for this problem:

[BI1] John M. McNamara and Par Forslund. Divorce rates in birds: Predictions from an optimization model. *The American Naturalist*, 147(4):609–640, 1996.

Chapter 5

Long run average payoff problems

5.1 Motivation

Consider a stationary Markov Decision Process $(X_k)_{k\geq 0}$, that is a MDP with a dynamics (transition probabilities) independent of time. This means that the following parameters (independent of time) are given:

- a state space $\mathcal{E}$, which is a finite or countable set;
- a set of actions $\mathcal{C}$, and possibly for each state $x \in \mathcal{E}$, a nonempty subset $\mathcal{C}(x)$ of $\mathcal{C}$, which is the set of possible actions when the state is equal to x;
- an initial probability $p^{(0)}$ on $\mathcal{E}$, or an initial state x_0, which is equivalent to the case where $p^{(0)}$ is the Dirac measure at x_0;
- for all $x \in \mathcal{E}$ and $u \in \mathcal{C}(x)$, a probability vector $M_x^{(u)}$ over $\mathcal{E}$, the entries of which will be denoted $\left(M_{xy}^{(u)}\right)_{y\in\mathcal{E}}$.

Recall, that given the above parameters, and a pure strategy $\sigma = (\sigma_k)_{k\geq 0}$, there exists two discrete time processes $(X_k)_{k\geq 0}$ and $(U_k)_{k\geq 0}$ taking their values in $\mathcal{E}$ and $\mathcal{C}$ respectively, with transition probabilities $M_{xy}^{(u)}$:

$$M_{xy}^{(u)} = P(X_{k+1} = y \mid X_k = x, U_k = u)$$

satisfying

$$U_k = \sigma_k(X_0, U_0, \ldots, X_{k-1}, U_{k-1}, X_k) \ ,$$

and the Markov property:

$$\begin{aligned} &P(X_{k+1} = y \mid X_k = x, U_k = u, X_{k-1} = x_{k-1}, U_{k-1} = u_{k-1}, \ldots, X_0 = x_0, U_0 = u_0) \\ &\quad = \quad P(X_{k+1} = y \mid X_k = x, U_k = u) \ , \quad \forall x, y, x_i \in \mathcal{E},\ u \in \mathcal{C}(x),\ u_i \in \mathcal{C}(x_i),\ \text{for } i \geq 0 \ . \end{aligned}$$

Moreover, the same holds for a relaxed strategy, in which case:

$$P(U_k \in B \mid X_0, U_0, \ldots, X_{k-1}, U_{k-1}, X_k) = [\sigma_k(X_0, U_0, \ldots, X_{k-1}, U_{k-1}, X_k)](B) \ .$$

We are interested here in the optimization of infinite horizon undiscounted criteria or in long run average criteria, for which we look for optimal strategies (among all strategies) that would be stationary and in feedback form (Markov).

We assume given the stationary *instantaneous/running reward/payoff* (at any time k), which is a map $r : \mathcal{A} \to \mathbb{R}$.

For all strategies $\sigma = (\sigma_k)_{k\geq 0}$ in Σ or Σ^{R}, the *undiscounted total additive payoff* with infinite horizon (when it exists) is:

$$J^{(\infty,\sigma)} := J^{\infty}(X;U) := \mathbb{E}\left[\sum_{k=0}^{\infty} r(X_k, U_k)\right] \ , \tag{5.1}$$

where $(X, U) := (X_k, U_k)_{k\geq 0}$ is the process induced by σ as in Definition 3.2 or Definition 3.3.

This criteria is finite when for instance there exist a cemetery point c in which the state process arrives almost surely in a finite expected time and the reward there is zero. In that case, one can often transform the problem in a discounted control problem, in which the discount factor α is such that the probability to arrive in one step to the cemetery point is at least $1 - \alpha$. So we can generally apply the methods of the first part of the course. In particular, the value function

$$v^{\infty}(x) := \max J^{(\infty,\sigma)} \quad \text{when } X_0 = x \ ,$$

is equal to the limit of the value function v^T of the problem with finite horizon and (undiscounted) additive criteria:

$$v^T(x) = \max J^{(T,\sigma)}(x) \ , \tag{5.2}$$

with

$$J^{(T,\sigma)}(x) = \mathbb{E}\left[\sum_{k=0}^{T} r(X_k, U_k) \mid X_0 = x\right] \ . \tag{5.3}$$

The second type of criteria is the *mean-payoff* or *long run time average payoff/reward*, which is one of the following ones, for all strategies $\sigma = (\sigma_k)_{k\geq 0}$ in Σ or Σ^{R}

$$J^{(+,\sigma)} := J^{+}(X;U) := \limsup_{T\to\infty} \left\{ \frac{1}{T}\mathbb{E}\left[\sum_{k=0}^{T} r(X_k, U_k)\right]\right\}, \tag{5.4a}$$

$$J^{(-,\sigma)} := J^{-}(X;U) := \liminf_{T\to\infty} \left\{ \frac{1}{T}\mathbb{E}\left[\sum_{k=0}^{T} r(X_k, U_k)\right]\right\}, \tag{5.4b}$$

where $(X, U) := (X_k, U_k)_{k\geq 0}$ is the process induced by σ as in Definition 3.2 or Definition 3.3.

The corresponding value functions will be denoted:

$$\zeta^{\pm}(x) = \max J^{(\pm,\sigma)} \quad \text{when } X_0 = x \ .$$

One may try to understand in which situations both criteria are equal and independent of the initial law of the MDP, and to compare the limit of $(1 - \alpha)J^{\sigma}_{\alpha}$ when α goes to 1 from below. Moreover, we shall compare the optimum of the limit with the limit of the optimum.

We also look for an optimal strategy which is a feedback stationary policy.

We first study the uncontrolled case.

5.2 Long term behavior of Markov chains

5.2.1 Ergodicity of Markov chains

In this section, we shall study the uncontrolled case.

We assume that $(X_k)_{k\geq 0}$ is a stationary Markov chain on the finite state space $\mathcal{E}$ (for instance $\mathcal{E} = [n]$) and we denote by M its transition probability matrix. Then, if $X_0 = x$, we have

$$\mathbb{E}\left[\sum_{k=0}^{T} r(X_k)\right] = \sum_{k=0}^{T} [M^k r]_x \ ,$$

and when X_0 is random with law $p^{(0)}$, we have

$$\mathbb{E}\left[\sum_{k=0}^{T} r(X_k)\right] = \sum_{k=0}^{T} p^{(0)} M^k r \ .$$

Hence, the value functions $\zeta^{\pm}$ defined in the previous section reduce to

$$\zeta^{\epsilon}(x) = J^{\epsilon}((X_k)_{k\geq 0}) := \begin{cases} \limsup_{T\to\infty} \left\{ \frac{1}{T} \sum_{k=0}^{T} [M^k r]_x \right\} & \text{if } \epsilon = + \\ \liminf_{T\to\infty} \left\{ \frac{1}{T} \sum_{k=0}^{T} [M^k r]_x \right\} & \text{if } \epsilon = - \ . \end{cases} \tag{5.5}$$

Moreover, if we consider the case of a Markov chain (X_k) with initial law $p^{(0)}$ (not necessarily equal to the Dirac measure in some state x_0), then we are looking for

$$\zeta^{\epsilon}(p^{(0)}) = J^{\epsilon}((X_k)_{k\geq 0}) := \begin{cases} \limsup_{T\to\infty} \left\{ \frac{1}{T} \sum_{k=0}^{T} (p^{(0)} M^k r) \right\} & \text{if } \epsilon = + \\ \liminf_{T\to\infty} \left\{ \frac{1}{T} \sum_{k=0}^{T} (p^{(0)} M^k r) \right\} & \text{if } \epsilon = - \ . \end{cases} \tag{5.6}$$

Example 5.1. If $(X_n)_{n\geq 0}$ is a sequence of independent random variables with values in $\mathcal{E}$, then it is in particular a Markov chain with transition matrix M such that all rows are equal to the probability vector of X_0, that is $p^{(0)}$. Then, $(r(X_n))_{n\geq 0}$ is a sequence of i.i.d. random variables with expectation equal to $p^{(0)}r$ and the law of large numbers shows that

$$\frac{1}{T} \sum_{k=0}^{T} r(X_k) \underset{T\to\infty}{\longrightarrow} p^{(0)} r \quad \text{a.s.}$$

Taking the expectation, we get that $\zeta^{\pm}(p^{(0)}) = p^{(0)} r$.

In order to generalize the law of large numbers, or at least the easier expectation version above, to a Markov chain, one need the following notion.

Definition 5.2. We say that $m \in \Delta_{\mathcal{E}}$ is an invariant probability measure of the Markov chain $(X_n)_{n\geq 0}$ on $\mathcal{E}$ with transition matrix M, or simply of the matrix M, if m satisfies $mM = m$.

As an example, if X_n are i.i.d with laws p, then X_n is a Markov chain with invariant probability measure p.

Fact 5.3. If the initial law of a Markov chain is an invariant probability measure, $p^{(0)} = m$, then the law of X_n is equal to m for all $n \geq 0$.

In this case, it is easy to see that $\zeta^{\pm}(p^{(0)}) = mr$. This motivates the following definition.

Definition 5.4. We say that the Markov chain with Markov transition matrix M is *ergodic* if M has a unique invariant probability measure m.

To check the ergodicity of the chain or to find more generally the limit $\zeta^{\pm}(p^{(0)})$, we need to study the spectral properties of the Markov matrix M. To do this, we can use general linear algebra techniques in particular Jordan normal form and/or the Perron-Frobenius theorem which is the tool for studying matrices with nonnegative entries. The latter result uses the properties of the graph associated to M, defined in the following section.

5.2.2 Graph properties of a Markov matrix

Recall (see Definition 2.13) that to a nonnegative matrix M over $\mathcal{E}$, we associate a *digraph* denoted $\mathcal{G}(M)$, with set of nodes equal to $\mathcal{E}$ and set of arcs $\mathcal{A}$ the set of $(x, y) \in \mathcal{E} \times \mathcal{E}$ such that $M_{xy} > 0$.

Definition 5.5. Given a directed graph $\mathcal{G}$, with set of nodes $\mathcal{E}$, we define the relations on $\mathcal{E}$ such that for $x, y \in \mathcal{E}$,

- $x \to y$ if there exists a path from x to y of any length ≥ 0 in $\mathcal{G}$ (where a path of length 0 means that $x = y$).
- $x \sim y$ if $x \to y$ and $y \to x$.

Proposition 5.6. *For any digraph $\mathcal{G}$ with set of nodes $\mathcal{E}$, $\sim$ is an equivalence relation and $\to$ a preorder on $\mathcal{E}$. This defines a partition of $\mathcal{E}$ into equivalence classes for $\sim$, that are called* strongly connected components *of the graph $\mathcal{G}$. If $\mathcal{G} = \mathcal{G}(M)$, where M is a Markov matrix, they are also called* communication classes *of M. Moreover, the relation $\to$ becomes a partial order on the set $\mathcal{E}/\sim$ of strongly connected components of $\mathcal{G}$.*

Proof. $\to$ is reflexive, $x \to x$ $\forall x \in \mathcal{E}$, because paths of length 0 are allowed. It is transitive : $(x \to y$ and $y \to z \Rightarrow x \to z)$ $\forall x, y, z \in \mathcal{E}$, by concatenation of paths. So it is a preorder.

Therefore $\sim$ is also reflexive and transitive. Moreover, it is symmetric by definition: $(x \sim y \Leftrightarrow y \sim x)$ $\forall x, y \in \mathcal{E}$.

Recall that the equivalence class of x for $\sim$ is defined as $\bar{x} = \{y \in \mathcal{E}, x \sim y\}$, and that we have $\bar{x} = \bar{y}$ if $x \sim y$ and $\bar{x} \cap \bar{y} = \emptyset$ otherwise, so that equivalence classes define a partition of $\mathcal{E}$.

Since $x \sim y$ if $x \to y$ and $y \to x$, we get that $\to$ can be defined on the quotient set $\mathcal{E}/\sim$, that is the set of equivalence classes for $\sim$, and that on this quotient set, we have ($\bar{x} \to \bar{y}$ and $\bar{y} \to \bar{x}$) if and only if $\bar{x} = \bar{y}$, so that $\to$ becomes a (partial) order. □

Definition 5.7.

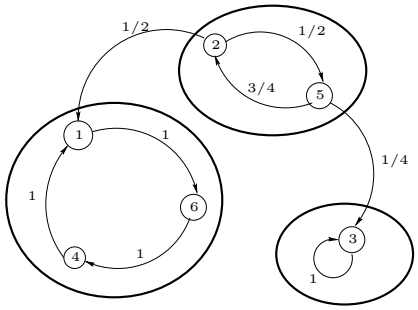
- A subset F of $\mathcal{E}$ is *closed* if $(x \in F$ and $x \to y) \Rightarrow y \in F$.
- A closed set with a unique element is *absorbant*.
- A maximal element for $\to$ in $\mathcal{E}/\sim$ is called a *final class*. This is a stongly connected component of $\mathcal{G}$ which is also a closed set.
- A *transient* state is an element of $\mathcal{E}$ which is not in a final class.
- The graph $\mathcal{G}$ is *strongly connected* if it has a unique strongly connected component, or equivalently if $\mathcal{E}$ is the only closed set.

- The matrix M is *irreducible* if $\mathcal{G}(M)$ is strongly connected.
- The Markov chain $(X_n)_{n\geq 0}$ is *irreducible* if its transition matrix is irreducible.

In a finite set $\mathcal{E}$, final classes are equivalent to *recurrence classes*.

Example 5.8. The digraph $\mathcal{G}(M)$ of the following matrix M is shown on the right:

$$M = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 1 \\ \frac{1}{2} & 0 & 0 & 0 & \frac{1}{2} & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{3}{4} & \frac{1}{4} & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix} \begin{matrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{matrix}.$$



We see 3 strongly connected components: $\{1,4,6\}$, $\{3\}$ and $\{2,5\}$. The state 3 is absorbant. The closed sets are $\mathcal{E}$, $\{3\}$, $\{1,4,6\}$ and $\{1,3,4,6\}$. The final classes are $\{3\}$ and $\{1,4,6\}$.

Example 5.9 (Simple random walk)**.** Given a directed graph $\mathcal{G}$ with set of nodes $\mathcal{N}$ and arcs $\mathcal{A}$, and at least one arc from each node, a simple random walk is a Markov chain with transition probabilities: $M_{xy} = 1/N_x$ if (x,y) is an arc and $N_x \geq 1$ is the number of arcs starting from x, and $M_{xy} = 0$ otherwise. This means that the random walker is going with uniform probability in outgoing arcs. The graph of M coincides with $\mathcal{G}$.

Example 5.10 (PageRank)**.** Denote by $\mathcal{E}$ the set of *Web pages*. The *graph of the Web* is composed of $\mathcal{E}$ as set of nodes, and contains an arc (x,y) if there is a *hyperlink* from page x to page y.

Assume that there is at least one hyperlink starting from any page. Otherwise, if a page x has no successor, then one add an arc from x to any page.

Let $P \in \mathbb{R}^{\mathcal{E}\times\mathcal{E}}$ be the Markov transition matrix of a simple random walk on the Web graph: $P_{xy} = 1/N_x$ if there is a hyperlink from x to y where $N_x \geq 1$ is the number of hyperlinks from x, and $P_{xy} = 0$ otherwise. This matrix may not be irreducible.

Google constructs the following Markov matrix:

$$M = \gamma P + (1-\gamma)\mathbf{1}z$$

where

- $0 < \gamma < 1$ is the *damping factor*: $1-\gamma$ is the probability that a Web surfer is stopping clicking on following pages and is returning to the Google site (or any search engine) for instance or is going to any page randomly;
- $z \in \Delta_{\mathcal{E}}$ is a row probability vector with positive entries ($z_x > 0$ for all $x \in \mathcal{E}$, and $z\mathbf{1} = 1$) giving the probability for the Web surfer of going to any page when he is stopping clicking on following pages. This is the *preference vector*.

The *PageRank* computed by Google [EC1] is the invariant measure p^M of M: that is a row probability vector ($p^M \in \Delta_{\mathcal{E}}$) satisfying $p^M = p^M M$. Hence, the PageRank of a Web page x, P_x^M,

corresponds to the expected frequency of visit of this page by the state of the Markov chain. Then the *order* on Web pages is defined as follows: a page x is *better* than page y if $p_x^M \geq p_y^M$.

In this context, a Website means a subset W of $\mathcal{E}$. To "optimize" his Website, the owner of W would like to maximize a certain positive linear combination of the PageRank of W:

$$\max_{P \in \mathcal{P}} \sum_{x \in W} g(x) p_x^M \quad , \tag{5.7}$$

where $g \in \mathbb{R}_+^W$, and $\mathcal{P}$ is a set of possible Markov matrices. To solve this problem, we shall interpret

$$\sum_{x \in W} g(x) p_x^M$$

as the value of a mean-payoff criteria, and in some particular cases the optimization as a Markov decision problem with mean-payoff criteria.

5.2.3 Perron-Frobenius theorem for irreducible matrices

The elements of this section can be found in [EC4].

Recall that we denote by $\leq$ the partial order on $\mathbb{R}^{\mathcal{E}}$ defined by entrywise inequalities: $v \leq w$ if $v_x \leq w_x$ for all $x \in \mathcal{E}$; that $\mathbf{1}$ is the column vector of $\mathbb{R}^{\mathcal{E}}$ with all its entries equal to 1: $\mathbf{1}_x = 1$ for all $x \in \mathcal{E}$; and that we denote by $\|\cdot\|_\infty$ the sup-norm on vectors and the associated norm on matrices. Then, any Markov Matrix M over a finite set $\mathcal{E}$ satisfies $\rho(M) = \|M\|_\infty = 1$ (see Lemma 2.26).

We also denote by $<$ the relation on $\mathbb{R}^{\mathcal{E}}$ defined by entrywise strict inequalities: $v < w$ if $v_x < w_x$ for all $x \in \mathcal{E}$.

Theorem 5.11 (Perron-Frobenius theorem)**.** *Let M be a matrix over $\mathcal{E}$ with nonnegative entries. Assume that M is irreducible. The following hold*

1. *$\rho(M)$ is an eigenvalue associated to an eigenvector $v_0 \in \mathbb{R}^{\mathcal{E}}$ with positive entries ($v_0 > 0$).*
2. *The eigenvalue $\rho(M)$ is (geomerically) simple, meaning that if $Mv = \rho(M)v$, with $v \in \mathbb{C}^{\mathcal{E}}$, then $v = \mu v_0$ for some scalar $\mu \in \mathbb{C}$.*
3. *Any eigenvector $v \geq 0$ is necessarily associated to the eigenvalue $\rho(M)$, and thus proportional to v_0.*
4. *If $Mv \leq \mu v$ and $v > 0$, then $\mu \geq \rho(M)$.* (Collatz-Wielandt property)
5. *If $\mu v \leq Mv$ with $\mu \geq 0$, $v \geq 0$ and $v \neq 0$, then $\mu \leq \rho(M)$.*
6. *If $Mv \leq \rho(M)v$ with $v \in \mathbb{R}^{\mathcal{E}}$ (or $Mv \geq \rho(M)v$), then $Mv = \rho(M)v$.*

Then, any eigenvector $v_0 > 0$ associated to the eigenvalue $\rho(M)$ is called a Perron vector*.*

Before giving the proof of Perron-Frobenius theorem, let us give some remark. Denote

$$\rho^+(M) = \inf\{\mu > 0 \mid \exists v > 0, Mv \leq \mu v\} = \inf_{v>0} \max_{x \in \mathcal{E}} \frac{\sum_{y \in \mathcal{E}} M_{xy} v_y}{v_x} \quad . \tag{5.8}$$

Note that this scalar is finite $< +\infty$ by the second equivalent formula. Point 4 of Perron-Frobenius theorem 5.11 says that $\rho^+(M) \geq \rho(M)$. Moreover, together with Point 1, it implies that the minimum is attained in the formula of $\rho^+(M)$ and that $\rho(M) = \rho^+(M)$.

Proof of Perron-Frobenius theorem. The proof consists in several steps that are not necessarily in the same order as the points in the theorem. Let n be the cardinality of $\mathcal{E}$ and assume that $\mathcal{E} = \{1, \dots, n\}$.

(1) If $Mv \leq \mu v$, with $v \geq 0$ and $v \neq 0$, then $v > 0$ and $\mu > 0$.
Indeed, since $M \geq 0$ is irreducible, then $(I+M)^n$ has positive entries. If $Mv \leq \mu v$, with $v \geq 0$ and $v \neq 0$, then $\mu \geq 0$ (applying $0 \leq [Mv]_x \leq \mu v_x$ to x such that $v_x \neq 0$), and so $(I+M)^n v \leq (1+\mu)^n v$ (by using the monotonicity of M). Therefore, $v > 0$ and so $\mu > 0$.

(2) Point 4 or equivalently $\rho^+(M) \geq \rho(M)$.
Indeed, let $v > 0$ such that $Mv \leq \mu v$. Considering the diagonal matrix D such that $D_{xx} = v_x$ for all $x \in \mathcal{E}$, we get that $D\mathbf{1} = v$, so $MD\mathbf{1} \leq \mu D\mathbf{1}$. Since $v > 0$, D is invertible and nonnegative, so $D^{-1}MD\mathbf{1} \leq \mu\mathbf{1}$. This implies that $\|D^{-1}MD\|_\infty \leq \mu$, by the above formula for the sup-norm, and so $\rho(M) = \rho(D^{-1}MD) \leq \mu$, which shows Point 4.

(3) The infinum in (5.8) is a minimum.
By definition of $\rho^+(M)$, there exists $\mu_n > \rho^+(M)$ and $v_n > 0$ such that $\lim_{n\to\infty} \mu_n = \rho^+(M)$ and $\mu_n v_n \geq M v_n$. One can choose v_n such that $v_n \cdot \mathbf{1} = 1$. Then, the sequence v_n is bounded and thus admits a converging subsequence, that we also denote v_n. Let v be the limit of this sequence. We have $v \geq 0$ and $v \neq 0$ and $Mv \leq \rho^+(M)v$. By Property (1) above, this implies that $v > 0$ and so $\rho^+(M)$ is a minimum.

(4) If v is such that $Mv \leq \rho^+(M)v$ and $v \geq 0$ then $Mv = \rho^+(M)v$.
Denote $w = \rho^+(M)v - Mv$. We have $w \geq 0$. Assume by contradiction that $w \neq 0$. Then, applying $(I+M)^n$ to w, we get that $(I+M)^n w > 0$ and that $(I+M)^n w = \rho^+(M)z - Mz$, with $z = (I+M)^n v > 0$. Then, $Mz < \rho^+(M)z$ and so there exists $\mu < \rho^+(M)$ such that $Mz \leq \mu z$, which contradicts the definition of $\rho^+(M)$.

(5) Point 1.
By (3), there exists $v > 0$ such that $Mv \leq \rho^+(M)v$. By (4), this implies that $Mv = \rho^+(M)v$. Hence, $\rho^+(M) \leq \rho(M)$ and since the reverse inequality holds by (2), we get that $\rho^+(M) = \rho(M)$ and that there exists $v > 0$ such that $Mv = \rho(M)v$, which shows Point 1.

(6) Point 6.
Let $v \in \mathbb{R}^{\mathcal{E}}$ such that $Mv \leq \rho(M)v$. Considering v_0 as in Point 1. There exists $\mu \in \mathbb{R}$ such that $v \geq \mu v_0$. Take μ as large as possible so that $z = v - \mu v_0 \geq 0$ and there exists an entry of z equal to zero. We have $Mz \leq \rho(M)z$ and $z \geq 0$ so by (1), if $z \neq 0$, this implies that $z > 0$ a contradiction. So $z = 0$, then $v = \mu v_0$ and $Mv = \rho(M)v$.

(7) Point 2.
Let $v \in \mathbb{C}^{\mathcal{E}} \setminus \{0\}$ be such that $Mv = \rho(M)v$. Taking the absolute value of the entries, we get that $\rho(M)|v| = |Mv| \leq M|v|$. Then, using Point 6, we deduce that $\rho(M)|v| = M|v|$ and so $|v| = \mu v_0$ for some constant $\mu \geq 0$. We also have $(1+\rho(M))^n v = (I+M)^n v$ and so $(1+\rho(M))^n |v| = |(I+M)^n v| \leq (I+M)^n |v| = (1+\rho(M))^n |v|$. Hence, $|(I+M)^n v| = (I+M)^n |v|$, in particular denoting $\alpha_x = (I+M)^n_{1x}$, we get $|\sum_{x=1}^n \alpha_x v_x| = \sum_{x=1}^n \alpha_x |v_x|$. Since all the α_x are > 0, this shows that there exists $\beta \in \mathbb{C}$, such that $v_x = \beta |v_x|$. So $v = \beta\mu v_0$.

(8) Point 5.
Let $v \geq 0$ such that $\mu v \leq Mv$ and $v \neq 0$. Then, by the monotonicity of M, we have $\mu^n v \leq M^n v$. Taking the sup-norm, we obtain that $\mu \leq \|M^n\|_\infty^{1/n}$ for all $n \geq 1$. Taking the limit when n goes to infinity, we deduce that $\mu \leq \rho(M)$.

(9) Point 3.
If $v \geq 0$, $v \neq 0$ is such that $Mv = \mu v$, then $\mu > 0$ and $v > 0$, by (1). So by Points 4 and 5, we get

that $\mu = \rho(M)$. Point 2 or 6 implies that v is proportional to v_0.

□

Corollary 5.12. *Let M be an irreducible Markov matrix. Then $\mathbf{1}$ is the unique eigenvector associated to the eigenvalue 1, up to a scalar factor, and there exists a unique invariant probability measure, that is a row vector m over $\mathcal{E}$ such that $m \geq 0$, $m\mathbf{1} = 1$.*

Proof. Since $\rho(M) = 1$, $M\mathbf{1} = \mathbf{1}$, and $\mathbf{1} > 0$, Point 2 or 3 of Perron-Frobenius theorem 5.11 implies that $\mathbf{1}$ is a Perron vector and any eigenvector associated to the eigenvalue 1 is proportional to $\mathbf{1}$. Since the transpose matrix of M, denoted M^T, is also nonnegative and $\rho(M^T) = \rho(M)$, then by Point 1 of Perron-Frobenius theorem 5.11, there exists a column vector $w > 0$ such that $M^T w = w$. Choosing w such that $w \cdot \mathbf{1} = 1$, we get that $m = w^T$ is an invariant probability measure of M: $m\mathbf{1} = 1$ and $mM = m$. Moreover, by Point 2 or 3 of Perron-Frobenius theorem 5.11, if m and m' are both invariant probability measures, then since $m^\top$ and $(m')^\top$ are eigenvectors of M^T associated to the eigenvalue 1, they are proportional. Then using the condition $m\mathbf{1} = m'\mathbf{1} = 1$, we get that they $m = m'$. So the invariant probability measure of M is unique. □

Definition 5.13. We say that a Markov matrix M is *primitive* or *acyclic* if there exists $k \geq 1$ such that M^k is positive, meaning that all its entries are positive.

Proposition 5.14. *Let M be a primitive Markov matrix over $\mathcal{E}$. Then 1 is the unique eigenvalue with modulus 1.*

Proof. Let λ be an eigenvalue with modulus 1 and v be an eigenvector associated to the eigenvalue λ. For any vector w in $\mathbb{R}^{\mathcal{E}}$, we denote by $|w|$ the vector with entries $|w_x|$, $x \in \mathcal{E}$. Then, $|v| = |\lambda v| = |Mv| \leq M|v|$. By Point 6 of Theorem 5.11, this implies that $|v|$ is proportional to the vector $\mathbf{1}$ and that $|\lambda v| = M|v|$. Hence, $|M^k v| = |\lambda^k v| = M^k|v|$ for all $k \geq 0$. Let k be such that M^k is positive, and let $x \in \mathcal{E}$. Then, taking the equality $|M^k v| = M^k|v|$ at x, we get that $|\sum_{y\in\mathcal{E}}[M^k]_{xy}v_y| = \sum_{y\in\mathcal{E}}[M^k]_{xy}|v_y|$. This implies that all the entries v_y have same "sign", that v is proportional to $|v|$ and so to $\mathbf{1}$. Hence, $\lambda = 1$. □

5.2.4 Linear algebra techniques and the multichain case

Proposition 5.15. *Let M be a Markov matrix M over the state space $\mathcal{E}$. Then all its eigenvalues of modulus 1 are semi-simple, meaning that they have no nilpotent.*

Proof. Let $M = QJQ^{-1}$ be the Jordan decomposition of M. Since $\|M\|_\infty = 1$, and $J^n = Q^{-1}M^nQ$, we deduce that $\|J^n\|_\infty \leq C = \|Q^{-1}\|_\infty\|Q\|_\infty$. If J' is a block of J of size k corresponding to an eigenvalue λ of modulus 1, then $J' = \lambda I + N$ where I is the identity matrix, and N is the nilpotent matrix of order k ($N^k = 0$) of the form $N = \begin{bmatrix} 0 & 1 & 0 & \cdots \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ \cdots & & & 0 \end{bmatrix}$. Then, $(J')^n = \sum_{i=0}^{k-1}\binom{n}{i}\lambda^{n-i}N^i$ (with $N^0 = I$) and $\|J^n\|_\infty \geq \|(J')^n\|_\infty \geq \binom{n}{k-1} - \sum_{i=0}^{k-2}\binom{n}{i}$. If $k > 1$, we get that $\|J^n\|_\infty$ tends to $+\infty$ when n goes to infinity, a contradiction. So $k = 1$ and the eigenvalue λ of M has no nilpotent. □

Corollary 5.16. *Let M be an irreducible Markov matrix. Then the eigenvalue 1 is algebraically simple.*

Proof. By Corollary 5.12, there is a unique eigenvector associated to the eigenvalue 1, so 1 is geometrically simple. From Proposition 5.15, 1 has no nilpotent, so 1 is algebraically simple. □

Proposition 5.17. *Given a Markov matrix M over the state space $\mathcal{E}$, and a Markov chain $(X_n)_{n\geq 0}$ with transition matrix M and initial state $X_0 = x$, we have*

$$\zeta^{\pm}(x) = [Pr]_x, \tag{5.9}$$

where P is the spectral projector of M for the eigenvalue 1, that is P is the unique matrix such that

$$P = P^2, \quad \operatorname{Im} P = \ker(I - M), \quad \ker P = \operatorname{Im}(I - M), \quad \textit{and} \quad P = PM = MP \ .$$

This implies in particular, in the uncontrolled case, that $\frac{v^T}{T}$ converges towards $\zeta^{\pm}$.

Proof. Recall that $\zeta^{\pm}(x)$ is the limsup or liminf of $\frac{1}{T}\sum_{k=0}^{T}[M^k r]_x$ when T goes to infinity. Let $M = QJQ^{-1}$ be the Jordan decomposition of M. Then, $\frac{1}{T}\sum_{k=0}^{T} M^k = Q\left(\frac{1}{T}\sum_{k=0}^{T} J^k\right)Q^{-1}$. Let J' be a block of J corresponding to an eigenvalue λ. If $|\lambda| < 1$, then $(J')^T$ tends to 0 when T goes to infinity, so does the Cesàro mean $\frac{1}{T}\sum_{k=0}^{T} J^k$. If $|\lambda| = 1$, then by Proposition 5.15, J' is a block of size 1 with entry λ. So $\frac{1}{T}\sum_{k=0}^{T} J^k$ is a block of size 1 and entry $\frac{1}{T}\sum_{k=0}^{T}\lambda^k$. If $\lambda \neq 1$, then this entry is equal to $(1-\lambda^{T+1})/(1-\lambda)/T$ which tends to 0 when T goes to infinity. Otherwise, the entry is equal to $1 + 1/T$ which tends to 1 when T goes to infinity. All together, we get that $\frac{1}{T}\sum_{k=0}^{T} J^k$ tends to the diagonal matrix D with ones at the places corresponding to the blocks of J associated to the eigenvalue 1 and 0 elsewhere. This diagonal matrix is exactly the spectral projector of J for the eigenvalue 1. Then, $\frac{1}{T}\sum_{k=0}^{T} M^k$ tends to QDQ^{-1}, which is the spectral projector P of M for the eigenvalue 1. Since 1 has no nilpotent, then P satisfies $\operatorname{Im} P = \ker(I-M)$ and $\ker P = \operatorname{Im}(I-M)$. □

Corollary 5.18. *Let M be an irreducible Markov matrix over the state space $\mathcal{E}$, let m be its unique invariant probability measure and let $(X_n)_{n\geq 0}$ be a Markov chain with transition matrix M and initial state $X_0 = x$. We have*

$$\zeta^{\pm}(x) = mr, \quad \forall x \in \mathcal{E} \ . \tag{5.10}$$

Proof. Let $M = QJQ^{-1}$ be the Jordan decomposition of M. Since the eigenvalue 1 of M is simple, the spectral projector P of M for the eigenvalue 1 is equal to QDQ^{-1}, where D is the diagonal matrix with 1 in some position x and 0 elsewhere. So P is the product of the column x of Q and of the row x of Q^{-1}, which are respectively equal to a column and row eigenvector of M with respect to the eigenvalue 1. These vectors are equal respectively to $\lambda\mathbf{1}$ and μm, for some $\lambda, \mu \in \mathbb{C}\setminus\{0\}$. Since $Q^{-1}Q = I$, and $m\mathbf{1} = 1$, we get that $\lambda\mu = 1$. So $P = \mathbf{1}m$ and the result follows. □

Theorem 5.19 (Decomposition of the spectral projector using final classes)**.** *Let M be a Markov matrix over the state space $\mathcal{E}$. For each final class $F \subset \mathcal{E}$, there exists a unique invariant probability measure $m^{(F)}$ of M with support equal to F, and a unique fixed point $v^{(F)}$ of M (that si satisfying $Mv^{(F)} = v^{(F)}$) such that $[v^{(F)}]_x = 1$ for $x \in F$ and $[v^{(F)}]_x = 0$ for $x \in F'$ and F' a final class $\neq F$. Moreover, the spectral projector of M for the eigenvalue 1 is equal to:*

$$P = \sum_{F \text{ final class}} v^{(F)} m^{(F)} \ .$$

Therefore, all invariant probability measures of M are convex combinations of the $m^{(F)}$, and all fixed points of M are linear combinations of the $v^{(F)}$. Given a Markov chain $(X_n)_{n\geq 0}$ with transition

matrix M and initial state $X_0 = x$, we have

$$\zeta^{\pm}(x) = \sum_{F \text{ final class}} (m^{(F)} r)[v^{(F)}]_x \ . \tag{5.11}$$

In particular

$$\zeta^{\pm}(x) = m^{(F)} r \quad \forall x \in F \ .$$

Proof. Let T be the set of transient states of M, that is the complementary of the union of final classes. Ordering the elements of $\mathcal{E}$ as $x_1, \ldots, x_n$ such that $x_i \to x_j$ for $i < j$, we get that the matrix M can be written in the following block form, where $F_1, \ldots, F_m$ are the final classes:

$$M = \begin{bmatrix} M_{TT} & M_{TF_1} & \cdots & M_{TF_m} \\ 0 & M_{F_1F_1} & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & M_{F_mF_m} \end{bmatrix} \ .$$

For each final class $F \in \{F_1, \ldots, F_m\}$, M_{FF} is an irreducible Markov matrix on the set of states F, so that it has a unique invariant probability measure m_F on F. For each final class, consider the row vector $m^{(F)} = \begin{bmatrix} 0 & \cdots & 0 & m_F & 0 & \cdots & 0 \end{bmatrix}$, where m_F corresponds to the restriction of $m^{(F)}$ to the states in F. It is easy to see that $m^{(F)}$ is an invariant probability measure of M with support equal to F. Conversely, if m is an invariant probability measure of M with support in F, then its restriction to F is an invariant probability measure of M_{FF} so it is equal to m_F and $m = m^{(F)}$.

The set T is equal to the union of the classes T_i, $i = 1, \ldots, k$, of M that are transient (that is not final). Then, the eigenvalues of M_{TT} are obtained by taking all the eigenvalues of the blocks $M_{T_iT_i}$ of M_{TT} corresponding to the classes T_i. Since T_i is a transient class of M, we have that $M_{T_iT_i}\mathbf{1} \leq \mathbf{1}$ and $M_{T_iT_i}\mathbf{1} \neq \mathbf{1}$. Using the irreducibility of $M_{T_iT_i}$ and applying Point 6 of Perron-Frobenius Theorem 5.11, we deduce that $\rho(M_{T_iT_i}) < 1$. Then, $\rho(M_{TT}) = \max_{i=1,\ldots,k} \rho(M_{T_iT_i}) < 1$.

Consider now the vector v_F on $T^c = F_1 \cup \cdots \cup F_m$ with entries $[v_F]_x$ equal to 1 for $x \in F$ and to 0 otherwise. Then, v_F is a fixed point of the restriction $M_{T^cT^c}$ of M to T^c. Consider $v^{(F)} = \begin{bmatrix} (I_{TT} - M_{TT})^{-1} M_{TF} v_F \\ v_F \end{bmatrix}$, which exists and has nonnegative entries, since $\rho(M_{TT}) < 1$. We have that $Mv^{(F)} = v^{(F)}$ and that the entries $[v^{(F)}]_x$ are equal to 1 for $x \in F$ and to 0 for $x \in F'$ with F' a final class $\neq F$. Conversely, if v satisfies these properties, then $v = v^{(F)}$. This finish the proof of the first assertion of the theorem.

In view of the block representation of M and of the properties that $\rho(M_{TT}) < 1$ and that the matrices $M_{F_jF_j}$ are irreducible and Markov, we get that 1 is an eigenvalue of (geometric and algebraic) multiplicity m of M. We have already found m left eigenvectors of M, $m^{(F_j)}$, $j = 1, \ldots, m$ and m right eigenvectors of M, $v^{(F_j)}$, $j = 1, \ldots, m$. Moreover, $m^{(F_j)} v^{(F_k)} = \delta_{jk}$, so that one can construct a Jordan decomposition of M, $M = QJQ^{-1}$, such that the m first diagonal entries of J are ones, the m first columns of Q are the eigenvectors $v^{(F_j)}$, $j = 1, \ldots, m$, and the m first rows of Q^{-1} are the invariant probability measures $m^{(F_j)}$, $j = 1, \ldots, m$. Then, the spectral projector is equal to $P = QDQ^{-1}$, where D contains the Jordan blocks corresponding to the eigenvalue 1 of M, so is the diagonal matrix with its first m diagonal entries equal to 1 and remaining ones equal to 0. This leads to $P = \sum_{j=1,\ldots,m} v^{(F_j)} m^{(F_j)}$, that is the formula of P in the theorem. The last assertions follow from this formula. □

The following result gives a characterization of ergodicity of a Markov chain.

Corollary 5.20. *Let M be a Markov matrix over the state space $\mathcal{E}$. Then, M has a unique final class $F \subset \mathcal{E}$ if and only if there exists a unique invariant probability measure m of M (that is the associated Markov chain is ergodic). In that case, F is the support of m and $\mathbf{1}$ is the unique fixed point of M. Moreover the spectral projector of M for the eigenvalue 1 is equal to $P = \mathbf{1}m$. Therefore, given a Markov chain $(X_n)_{n\geq 0}$ with transition matrix M and initial state $X_0 = x$, we have*

$$\zeta^{\pm}(x) = mr \quad \forall x \in \mathcal{E} \ .$$

The previous results show that Cesàro means of M^n converge. Using Proposition 5.14, one shows that, in the primitive case, the following stronger property holds.

Proposition 5.21. *Let M be a primitive Markov matrix, and let ρ_2 be the maximum of the modulus of the eigenvalues $\neq 1$, and let k be the maximal size of the Jordan block of such an eigenvalue. We have*

$$M^n = \mathbf{1}m + \mathcal{O}(\rho_2^n n^{k-1}) \ .$$

The previous results, in particular Corollary 5.20 can be seen as a weak version of the ergodic theorem, stating a convergence in law. The following "strong" ergodic theorem states almost sure convergence. It can be proved using probabilistic techniques, and in particular using the law of large numbers, whereas it generalizes the law of large numbers. We state it without proof.

Theorem 5.22 (See []). *Let M and m be as in Corollary 5.20. Given a Markov chain $(X_n)_{n\geq 0}$ with transition matrix M, and any initial law $p^{(0)}$, we have*

$$\lim_{T\to\infty} \left\{ \frac{1}{T} \sum_{k=0}^{T} r(X_k) \right\} = mr, \qquad \textit{almost surely.} \tag{5.12}$$

5.2.5 The ergodic Kolmogorov equation

Consider first the case where M has a unique final (or recurrence) class. By Corollary 5.20, M has a unique invariant probability m, the support of m is the final class of M, and any right eigenvector of M associated to the eigenvalue 1 is a constant vector ($Mv = v \implies v = \lambda\mathbf{1}$ for some $\lambda \in \mathbb{C}$). This implies that $\zeta^{\pm}(x)$ is independent of the initial state $x \in \mathcal{E}$, and equal to $mr = \sum_{x\in\mathcal{E}} m_x r(x)$.

We can also characterize the value function $\zeta^{\pm}$ as follows.

Proposition 5.23 (The ergodic Kolmogorov equation)**.** *Let $\mathcal{E}$ be a finite set, M be a Markov transition matrix on $\mathcal{E}$ and $r \in \mathbb{R}^{\mathcal{E}}$. The following assertions hold:*

1. *Assume that there exists $\rho \in \mathbb{R}$ and $v \in \mathbb{R}^{\mathcal{E}}$ such that*

$$\rho\mathbf{1} + v = r + Mv \ . \tag{5.13}$$

   Then, $\zeta^{\pm}(x) = \rho$ for all $x \in \mathcal{E}$.

2. *Assume that M has a unique final class, then there exists $\rho \in \mathbb{R}$ and $v \in \mathbb{R}^{\mathcal{E}}$ satisfying* (5.13)*. Moreover, $\rho \in \mathbb{R}$ satisfying* (5.13) *is unique equal to mr, where m is the unique invariant probability measure of M and v satisfying* (5.13) *is unique up to an additive constant, meaning that if v, v' satisfy* (5.13)*, then $v - v'$ is a constant vector.*

Proof. 1) We already know that $\zeta^{\pm}(x) = [Pr]_x$, where P is the spectral projector of M for the eigenvalue 1. Then, applying P to (5.13), we get that $\rho P\mathbf{1} + Pv = Pr + PMv$ and since $P = PM$ and $P\mathbf{1} = \mathbf{1}$ ($\mathbf{1}$ is a right eigenvector of M), we obtain $\rho\mathbf{1} = Pr$, so $\zeta^{\pm}(x) = \rho$ for all $x \in \mathcal{E}$.

2) If M has a unique final class, then $P = \mathbf{1}m$, where m is the unique invariant probability measure of M, and so $\zeta^{\pm}(x) = [Pr]_x = mr$, for all $x \in \mathcal{E}$. Let $\rho = mr$, we get that $m(r - \rho\mathbf{1}) = 0$ so $r - \rho\mathbf{1} \in \ker m = \ker P = \mathrm{Im}(I - M)$. Hence, there exists $v \in \mathbb{R}^S$ such that $r - \rho\mathbf{1} = v - Mv$ that is ρ and v satisfy (5.13). Conversely, if ρ and v satisfy (5.13), then by applying m to the equation, we get that $\rho = mr$ so ρ is unique. Also if v, v' satisfy (5.13), then $(I - M)(v - v') = 0$ so $v - v'$ is a constant vector. □

Example 5.24 (Pagerank (continued))**.** Let us come back to Example 5.10. Recall that to "optimize" his Website, the owner of W would like to maximize the criteria (5.7).

Since M is irreducible (it has positive entries, since $z > 0$), we get that p^M is unique $p^M > 0$, and

$$\sum_{x \in W} g(x) p_x^M = p^M g = \rho$$

where g is extended by zero on W^c, $\rho = \zeta^{\pm}(x)$ for all $x \in \mathcal{E}$ with:

$$\zeta^{\epsilon}(x) = \begin{cases} \limsup_{T\to\infty} \left\{ \frac{1}{T}\mathbb{E}\left[\sum_{k=0}^{T} g(X_k) \mid X_0 = x\right]\right\} & \text{if } \epsilon = + \\ \liminf_{T\to\infty} \left\{ \frac{1}{T}\mathbb{E}\left[\sum_{k=0}^{T} g(X_k) \mid X_0 = x\right]\right\} & \text{if } \epsilon = - \end{cases} .$$

and there exists $v \in \mathbb{R}^{\mathcal{E}}$ satisfying the ergodic Kolmogorov equation:

$$\rho\mathbf{1} + v = g + Mv \ .$$

Moreover since $M = \gamma P + (1 - \gamma)\mathbf{1}z$, we have $v = g + \gamma Pv$ and $\rho = (1 - \gamma)zv$.

Proposition 5.23 can be generalized as follows. The word "multichain" refers to the case of Markov chains with multiple final/recurrence classes.

Proposition 5.25 (The multichain Kolmogorov equation)**.** *Let $\mathcal{E}$ be a finite set, M be a Markov transition matrix on $\mathcal{E}$ and $r \in \mathbb{R}^{\mathcal{E}}$, and let $\zeta = \zeta^{\pm} \in \mathbb{R}^{\mathcal{E}}$ be defined as in* (5.5) *or* (5.9)*. Then, there exists $v \in \mathbb{R}^{\mathcal{E}}$ such that (ζ, v) is solution to the following equations:*

$$\zeta + v = r + Mv \tag{5.14a}$$
$$\zeta = M\zeta \tag{5.14b}$$

The solution ζ of (5.14) *is unique and thus equal to Pr, and v is unique up to the addition of any element of* $\ker(I - M) = \ker(I - P)$*. In particular, there exists a unique (ζ, v) satisfying* (5.14) *together with the following condition:*

$$mv = 0 \quad \textit{for all invariant probability measures } m \textit{ of } M. \tag{5.15}$$

Note that (5.15) is equivalent to the condition that $Pv = 0$.

5.3 The controlled case

Let us consider now the controlled problem. Our first aim is to show a result similar to Proposition 5.23.

We assume given a stationary Markov decision process and a stationary *instantaneous/running reward/payoff* (at any time k), which is a map $r : \mathcal{A} \to \mathbb{R}$. For all strategies $\sigma = (\sigma_k)_{k\geq 0}$ in Σ or Σ^{R}, and initial state $x \in \mathcal{E}$, we consider the *mean-payoff* or *long run time average payoff/reward* :

$$J^{(+,\sigma)}(x) := J_x^+(X;U) := \limsup_{T\to\infty} \left\{ \frac{1}{T}\mathbb{E}\left[\sum_{k=0}^{T} r(X_k,U_k) \mid X_0 = x\right]\right\}, \tag{5.16a}$$

$$J^{(-,\sigma)}(x) := J_x^-(X;U) := \liminf_{T\to\infty} \left\{ \frac{1}{T}\mathbb{E}\left[\sum_{k=0}^{T} r(X_k,U_k) \mid X_0 = x\right]\right\}, \tag{5.16b}$$

where $(X,U) := (X_k,U_k)_{k\geq 0}$ is the process induced by σ as in Definition 3.2 or Definition 3.3.

Let $\mathcal{B} : \mathbb{R}^{\mathcal{E}} \to \mathbb{R}^{\mathcal{E}}$ be the Bellman dynamic programming operator associated to undiscounted Markov decision problem:

$$[\mathcal{B}(v)](x) = \sup_{u\in\mathcal{C}(x)} \left(r(x,u) + \sum_{y\in\mathcal{E}} M_{xy}^{(u)} v(y) \right) \ ,$$

for $v \in \mathbb{R}^{\mathcal{E}}$, and $x \in \mathcal{E}$.

For any feedback policy $\pi \in \Pi := \{\pi : \mathcal{E} \to \mathcal{C} \mid \pi(x) \in \mathcal{C}(x),\ \forall x \in \mathcal{E}\}$, we denote by $r^{(\pi)}$, $M^{(\pi)}$ and $\mathcal{B}^{(\pi)}$ the reward vector, Markov transition matrix and Kolmogorov operator of the Markov decision problem with fixed policy π:

$$r_x^{(\pi)} = r(x,\pi(x)) \ , \quad M^{(\pi)} = (M_{xy}^{(\pi(x))})_{x,y\in\mathcal{E}} \ , \quad \mathcal{B}^{(\pi)}(v) = r^{(\pi)} + M^{(\pi)} v \ .$$

5.3.1 The ergodic dynamic programming equation

Let us first assume

(A5) There exists $\rho \in \mathbb{R}$ and $v \in \mathbb{R}^{\mathcal{E}}$ satisfying the ergodic dynamic programming equation equation:

$$\rho\mathbf{1} + v = \mathcal{B}(v) \ . \tag{5.17}$$

Theorem 5.26 (The ergodic dynamic programming equation)**.** *Under* (A5)*, the value function of the mean-payoff (long run time average payoff) Markov decision problem :*

$$\zeta^{\pm}(x) := \max_{\sigma} J^{(\pm,\sigma)}(x) \ ,$$

where the maximum is taken over either all relaxed strategies (starting at time 0*), or over the restricted sets of pure strategies, Markov strategies, feedback policies, or stationary feedback policies, satisfies*

$$\zeta^{\pm}(x) = \rho \quad \forall x \in \mathcal{E} \ .$$

Moreover, if, for all $x \in \mathcal{E}$,

$$\pi(x) \in \operatorname*{Argmax}_{u \in \mathcal{C}(x)} \left(r(x,u) + \sum_{y \in \mathcal{E}} M_{xy}^{(u)} v(y) \right) \ ,$$

then π is an optimal stationary policy for the MDP with mean-payoff.

For the proof, we shall use the following result which computes the *limit of the supremum* instead of the *supremum of the limit*. Let v^T be the value of a finite horizon problem. From dynamic programming equation for MDP with finite horizon (Theorem 3.13) and the stationarity of the MDP and instantaneous reward, this is equivalent to compute recursively (forward): $v^T = \mathcal{B}(v^{T-1})$ with $v^0 = \varphi \in \mathbb{R}^{\mathcal{E}}$.

Proposition 5.27. *Assume that there exists $\rho \in \mathbb{R}$ and $v \in \mathbb{R}^{\mathcal{E}}$ satisfying the ergodic Bellman equation* (5.17)*. We have:*

1. $\lim_{T\to\infty} \frac{1}{T} v^T = \rho \mathbf{1}$
2. *$v^T - \rho T \mathbf{1}$ is bounded (w.r.t. $T > 0$).*

Proof. Let ρ and v satisfy $\rho\mathbf{1} + v = \mathcal{B}(v)$, and v^T satisfy $v^T = \mathcal{B}(v^{T-1})$ with $v^0 = \varphi$. Consider also the finite horizon value w^T starting from $w^0 = v$. Since $\mathcal{B}$ is additively homogeneous, and $\mathcal{B}(v) = \rho\mathbf{1} + v$, we obtain

$$w^T = \mathcal{B}^T(v) = T\rho\mathbf{1} + v \ .$$

Since $\mathcal{B}$ is Lipschitz continuous with constant 1 (one also says nonexpansive), we deduce:

$$\|w^T - v^T\|_\infty \leq \|w^0 - v^0\|_\infty = \|v - \varphi\|_\infty \ .$$

So

$$\|v^T - \rho T\mathbf{1}\|_\infty \leq \|v^T - w^T\|_\infty + \|v\|_\infty \leq \|v - \varphi\|_\infty + \|v\|_\infty \ .$$

This shows Point 2, which in turn implies Point 1. □

Proof of Theorem 5.26. **Proof of $\zeta^+ \leq \rho\mathbf{1}$.**

Let σ be any strategy. For all finite horizon T, denote

$$J^{(T,\sigma)}(x) = \mathbb{E}\left[\sum_{k=0}^{T} r(X_k, U_k) \mid X_0 = x \right] \ .$$

Then,

$$J^{(+,\sigma)}(x) = \limsup_{T\to\infty} \left\{ \frac{1}{T} J^{(T,\sigma)}(x) \right\} \ .$$

Let v^T be the value of the MDP with finite horizon and final reward $\varphi = 0$. Then,

$$J^{(T,\sigma)}(x) \leq v^{T+1}(x) \ .$$

Using Proposition 5.27, we get

$$J^{(+,\sigma)}(x) = \limsup_{T\to\infty} \left\{ \frac{1}{T} J^{(T,\sigma)}(x) \right\} \leq \lim_{T\to\infty} \left\{ \frac{1}{T} v^{T+1}(x) \right\} = \rho \ .$$

Since this holds for all strategies, we get $\zeta^+(x) \leq \rho$, where the supremum is taken over all relaxed strategies.

Proof of $\zeta^- \geq \rho\mathbf{1}$. Assume first that the maximum in the ergodic dynamic programming equation is attained by some policy $\pi \in \Pi$. Then ρ and v satisfy the ergodic Kolmogorov equation

$$\rho\mathbf{1} + v = r^{(\pi)} + M^{(\pi)}v = \mathcal{B}^{(\pi)}(v) \ .$$

Using Proposition 5.23, we deduce that

$$\rho = J^{(\pm,\pi)}(x) \quad \forall x \in \mathcal{E} \ .$$

Hence $\zeta^-(x) \geq J^{(-,\pi)}(x) = \rho$ for all $x \in \mathcal{E}$.

Assume now more generally that there exists $\pi \in \Pi$ which is ε-optimal in the r.h.s of the ergodic dynamic programming equation. Then,

$$-\varepsilon\mathbf{1} + \rho\mathbf{1} + v = -\varepsilon\mathbf{1} + \mathcal{B}(v) \leq r^{(\pi)} + M^{(\pi)}v \ .$$

Using the same technique as in the proof of Proposition 5.23, that is multiplying by the spectral projector $P^{(\pi)}$ of $M^{(\pi)}$, we obtain:

$$(\rho - \varepsilon)\mathbf{1} + P^{(\pi)}v \leq P^{(\pi)}r^{(\pi)} + P^{(\pi)}M^{(\pi)}v$$

then

$$(\rho - \varepsilon)\mathbf{1} \leq P^{(\pi)}r^{(\pi)} \ .$$

Since

$$[P^{(\pi)}r^{(\pi)}]_x = \lim_{T\to\infty} \left\{ \frac{1}{T}\mathbb{E}\left[\sum_{k=0}^{T} r(X_k, U_k) \mid X_0 = x \right] \right\} = J^{(-,\pi)}(x) \leq \zeta^-(x)$$

we obtain

$$\rho - \varepsilon \leq \zeta^-(x) \quad \forall x \in \mathcal{E} \ .$$

Since this holds for all $\varepsilon > 0$, we obtain $\zeta^- \geq \rho\mathbf{1}$. □

Proposition 5.27 suggests the following algorithm.

Definition 5.28. *Relative value iterations* consists in computing the sequence $v^T - v^T(x_0)\mathbf{1}$ for some fixed state x_0.

Example 5.29. In general, $v^T - v^T(x_0)\mathbf{1}$ or $v^T - \rho T\mathbf{1}$ does not converge when T goes to infinity. Let us show an example in the uncontrolled deterministic case. Consider the Bellman operator:

$$\mathcal{B}(v) = \begin{bmatrix} 1 \\ -1 \end{bmatrix} + \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} v$$

The Markov chain is a deterministic process: $1 \to 2$ and $2 \to 1$, and the reward satisfies $r(1) = 1$ and $r(2) = -1$. The invariant measure is $[1/2\ 1/2]$ and so $\rho = 0$. Starting from $v^0 = \varphi = 0$, we obtain

$$v^1 = \mathcal{B}(v^0) = \begin{bmatrix} 1 \\ -1 \end{bmatrix}, \quad v^2 = \mathcal{B}(v^1) = 0 = v^0 \ .$$

So $v^T - \rho T\mathbf{1}$ does not converge, nor $v^T - v^T(1)\mathbf{1}$:

$$v^T - v^T(1)\mathbf{1} = \begin{bmatrix} 0 \\ 2 \end{bmatrix} \text{ for } T \text{ odd}, \quad = 0 \text{ for } T \text{ even}.$$

Remark 5.30. One way to improve the method is to apply relative value iteration combined with Krasnoselskii-Mann iterations with respect to $\mathcal{B}$, that is compute $w^T - w^T(x_0)\mathbf{1}$, where

$$w^{T+1} = \frac{1}{2}w^T + \frac{1}{2}\mathcal{B}(w^T) \ .$$

Note that (ρ, v) satisfies ergodic equation if and only if

$$\frac{1}{2}\rho\mathbf{1} + v = \mathcal{B}^{\mathrm{KM}}(v) := \frac{1}{2}v + \frac{1}{2}\mathcal{B}(v) \ .$$

Moreover, if M is irreducible, then $(I + M)/2$ is primitive and so $((I + M)/2)^n$ converges, see Proposition 5.21.

5.3.2 Application to Pagerank optimization

Let us come back to Examples 5.10 and 5.24. The owner of a Web site W can choose any hyperlink starting from any page belonging to W according to his own rules, but cannot change the hyperkinks starting from the other Web pages.

Assume that the rules are such that any page in W *can be changed independently from the other pages*, so that the set $\mathcal{P}$ of possible Markov matrices P is of the form:

$$\mathcal{P} = \{P \in \mathbb{R}_+^{\mathcal{E}\times\mathcal{E}} \mid P_{x\cdot} \in \mathcal{C}(x)\}$$

where $P_{x\cdot}$ denotes the row x of P, $\mathcal{C}(x)$ is a subset of $\mathcal{C} := \Delta_{\mathcal{E}}$, and $\mathcal{C}(x)$ is a singleton when $x \notin W$.

Considering the *Markov decision process* with state space $\mathcal{E}$, action spaces $\mathcal{C}(x)$, $x \in \mathcal{E}$, and transition probabilities:

$$M_{xy}^{(u)} = \gamma u_y + (1-\gamma)z_y \ ,$$

we obtain that if $\pi \in \Pi$ is the policy such that $\pi(x) = P_{x\cdot}$, then

$$M_{xy}^{(\pi)} = M_{xy}^{(\pi(x))} = \gamma[\pi(x)]_y + (1-\gamma)z_y = \gamma P_{xy} + (1-\gamma)z_y = M_{xy} \ .$$

Conversely, if P is the matrix such that $P_{x\cdot} = \pi(x)$, then $M = M^{(\pi)}$.

The optimization rewrites as the *mean-payoff Markov decision problem*:

$$\max_{\pi\in\Pi} J^{(\pm,\pi)}(x') \ ,$$

for any $x' \in \mathcal{E}$, with

$$J^{(+,\pi)}(x) := J_x^+(X;U) := \limsup_{T\to\infty}\left\{\frac{1}{T}\mathbb{E}\left[\sum_{k=0}^{T} g(X_k) \mid X_0 = x\right]\right\},$$
$$J^{(-,\pi)}(x) := J_x^-(X;U) := \liminf_{T\to\infty}\left\{\frac{1}{T}\mathbb{E}\left[\sum_{k=0}^{T} g(X_k) \mid X_0 = x\right]\right\}.$$

where g is extended by 0 on $\mathcal{E}\setminus W$, $U_k = \pi(X_k)$ and $P(X_{k+1} = y \mid X_x = x, U_k = u) = M_{xy}^{(u)}$. An *optimal stationary feedback policy* for the MDP corresponds to an *optimal matrix* P for the optimization problem.

Consider the associated ergodic dynamic programming equation:

$$\rho \mathbf{1} + v = \mathcal{B}(v)$$

with

$$\begin{aligned}
[\mathcal{B}(v)]_x &= \max_{u \in \mathcal{C}(x)} \left\{ g(x) + \sum_{y \in \mathcal{E}} M_{xy}^{(u)} v(y) \right\} \\
&= \max_{u \in \mathcal{C}(x)} \left\{ g(x) + \sum_{y \in \mathcal{E}} (\gamma u_y + (1-\gamma) z_y) v(y) \right\} \\
&= g(x) + \gamma \max_{u \in \mathcal{C}(x)} \left\{ \sum_{y \in \mathcal{E}} u_y v(y) \right\} + (1-\gamma) \left\{ \sum_{y \in \mathcal{E}} z_y v(y) \right\}
\end{aligned}$$

It can be rewritten in the form:

$$\rho + v(x) = g(x) + \gamma \max_{u \in \mathcal{C}(x)} (u \cdot v) + (1-\gamma) z \cdot v \ .$$

Note that since $z \cdot v$ is a constant, then one can solve first:

$$v(x) = g(x) + \gamma \max_{u \in \mathcal{C}(x)} (u \cdot v) \ ; \tag{5.18}$$

and then take

$$\rho = (1-\gamma) z \cdot v \ .$$

(5.18) is the *dynamic programming equation of a discounted infinite horizon Markov decision problem* with discount factor γ, so it has a unique solution. This yields a solution to the ergodic equation. (5.18) can be solved, either by value iterations or by policy iterations.

Here are some examples or characteristics of the sets $\mathcal{C}(x)$:

- The owner of W cannot act on the pages not in W, so for $x \in \mathcal{E} \setminus W$, $\mathcal{C}(x)$ is a singleton of $\Delta_{\mathcal{E}}$.
- Without any constraint on the Web pages, except that P is the transition probability matrix of a simple random walk, $\mathcal{C}(x)$ is the set of uniform probabilities on nonempty subsets of $\mathcal{E}$. Note that the cardinality of $\mathcal{C}(x)$ is 2^N where N is the cardinality of $\mathcal{E}$, which lead to an exponential complexity of the solution of the problem. To avoid this exponential complexity, or to modelize some inequalities in hyperlinks due for instance to the order of the hyperlinks in the web page, or to the text size and font of the hyperlinks, one may choose to relax the problem by considering $\mathcal{C}(x) = \Delta_{\mathcal{E}}$.
- Let $\mathcal{F}$ be a set of *forbidden pages*, $\mathcal{M}$ a set of *mandatory pages*, and $\mathcal{R}$ the complement. Then, $\mathcal{C}(x)$ is the set of uniform probabilities on subsets A of $\mathcal{E}$ such that $\mathcal{M} \subset A \subset \mathcal{E} \setminus \mathcal{F}$. This can be relaxed by considering $\mathcal{C}(x)$ as the set of $p \in \Delta_{\mathcal{E}}$ such that $p_y \geq p_{y_0}$ for $y \in \mathcal{M}$ and $p_y = 0$ for $y \in \mathcal{F}$, where p_{y_0} is the uniform probability on $\mathcal{E} \setminus \mathcal{F}$.
- *Skeleton constraints*: one can consider the set of p such that $p_y \geq (1-\mu) q_y$ where $q \in \Delta_{\mathcal{E}}$.

- *Conditionnal probability constraints* $P(X_{k+1} \in J_x \mid X_k = x) \leq b$ is equivalent to

$$\mathcal{C}(x) = \{p \in \Delta_{\mathcal{E}} \mid \sum_{j \in J_x} p_j \leq b\} \ .$$

- *Frequency constraints* $P(X_{k+1} \in J \mid X_k \in I) \leq b$ cannot be put into a *local* constraint of the form $\mathcal{C}(x)$.

The problem can be solved analytically in the following particular examples.

- Assume that there is no hyperlink from Web pages outside W to Web pages inside W. Since $g(x) = 0$ for $x \notin W$, the solution of (5.18) satisfies

$$v(x) = \gamma \max_{u \in \mathcal{C}(x)} (u \cdot v), \quad \forall x \in W^c \ ,$$

  where $u \in \mathcal{C}(x)$ has a support in W^c. This equation depends only on $v|_{W^c}$, and it is a fixed point equation of a contracting map. Since it has 0 has a solution, we get $v|_{W^c} = 0$.

  The remaining equations reduce to equations for W states only:

$$v(x) = g(x) + \gamma \max_{u \in \mathcal{C}(x)} (\sum_{y \in W} u_y v(y)) \quad \forall x \in W \ .$$

- If $g(x) > 0$ on W, and $W = \{x_0\}$ is a single page, this reduces to

$$v(x_0) = g(x_0) + \gamma \max_{u \in \mathcal{C}(x_0)} (u_{x_0} v(x_0)) \ .$$

  So $v(x_0) > 0$ and the optimal u need to maximize u_{x_0}, that is the self hyperlink among the possible constraints.

- If $g(x) > 0$ on W and $\mathcal{C}(x) = \Delta_{\mathcal{E}}$, then

$$v(x) = g(x) + \gamma \max_{u \in \Delta_{\mathcal{E}}} \sum_{y \in W} (u_y v(y)) \ .$$

  so $v = g + \mu \mathbf{1}$ on W for some $\mu \in \mathbb{R}$. Moreover, the maximum in u is satisfied for u such that there exists $\lambda \in \mathbb{R}$ and $\lambda_y \geq 0$ (the Lagrange multipliers for the constraints $u\mathbf{1} = 1$ and $u_y \geq 0$ respectively) with $v(y) = \lambda - \lambda_y$ and $\lambda_y u_y = 0$ (complementary slackness). If g takes only different values on W, then, $u = \delta_{x_0}$ for $x_0 \in \operatorname{Argmax} g(x)$.

The initial optimization problem can also be generalized as follows. Consider the optimization problem

$$\max_{P \in \mathcal{P}} \sum_{x \in W} g(x, P_{x\cdot}) p_x^M \ ,$$

where now the criteria depends also on the rows of P. For instance if $g(x, u)$ is linear with respect to $\gamma u + (1 - \gamma) z$, we obtain:

$$\max_{P \in \mathcal{P}} \sum_{x,y \in \mathcal{E}} g_{xy} M_{xy} p_x^M \ ,$$

and $M_{xy}p_x^M$ represents the probability to move from Web page x to Web page y, when the invariant probability measure is applied, so in the long run.

This optimization problem rewrites as a *mean-payoff Markov decision problem*, with instantaneous reward g depending on u. It can be solved by computing ρ and v solutions of the ergodic equation:

$$\rho\mathbf{1} + v = \mathcal{B}(v)$$

with

$$\begin{aligned}[\mathcal{B}(v)]_x &= \max_{u\in\mathcal{C}(x)} \left\{ g(x,u) + \sum_{y\in\mathcal{E}} M_{xy}^{(u)} v(y) \right\} \\ &= \max_{u\in\mathcal{C}(x)} \left\{ g(x,u) + \gamma \sum_{y\in\mathcal{E}} u_y v(y) \right\} + (1-\gamma) \left\{ \sum_{y\in\mathcal{E}} z_y v(y) \right\} \quad ,\end{aligned}$$

which can be rewritten in the form:

$$\rho + v(x) = \max_{u\in\mathcal{C}(x)} (g(x,u) + \gamma u \cdot v) + (1-\gamma) z \cdot v \quad .$$

As above, one can solve first:

$$v(x) = \max_{u\in\mathcal{C}(x)} (g(x,u) + \gamma u \cdot v) \quad ; \tag{5.19}$$

and then take

$$\rho = (1-\gamma) z \cdot v \quad .$$

(5.19) is the *dynamic programming equation of a discounted infinite horizon Markov decision problem* with discount factor γ, so it has a unique solution. This yields a solution to the ergodic equation. Moreover, (5.19) can be solved, either by value iterations or by policy iterations.

5.3.3 Vanishing discount approach

Another way to improve relative value iterations is to replace the uniform mean in time by a discounted mean in time.

Let us first consider the uncontrolled case, that is consider a stationary Markov chain $(X_k)_{k\geq 0}$ on the finite state space $\mathcal{E}$, and denote by M its transition probability matrix, and by $r \in \mathbb{R}^{\mathcal{E}}$ a running reward vector.

The following criteria are similar to the mean-payoff criteria:

$$\begin{aligned}\zeta^+(x) := J^+((X_k)_{k\geq 0}) &= \limsup_{\alpha\to 1^-} \left\{ (1-\alpha)\mathbb{E}\left[\sum_{k=0}^{\infty} \alpha^k r(X_k) \mid X_0 = x \right] \right\}, \\ \zeta^-(x) := J^-((X_k)_{k\geq 0}) &= \liminf_{\alpha\to 1^-} \left\{ (1-\alpha)\mathbb{E}\left[\sum_{k=0}^{\infty} \alpha^k r(X_k) \mid X_0 = x \right] \right\},\end{aligned}$$

which can be rewritten as

$$\zeta^\epsilon(x) = \begin{cases} \limsup_{\alpha\to 1^-} \left\{ (1-\alpha) \sum_{k=0}^{\infty} [(\alpha M)^k r]_x \right\} & \text{if } \epsilon = + \\ \liminf_{\alpha\to 1^-} \left\{ (1-\alpha) \sum_{k=0}^{\infty} [(\alpha M)^k r]_x \right\} & \text{if } \epsilon = - \quad . \end{cases} \tag{5.20}$$

Proposition 5.31. *Given a Markov matrix M over the state space $\mathcal{E}$, and a Markov chain $(X_n)_{n\geq 0}$ with transition matrix M and initial state $X_0 = x$, the criteria in* (5.20) *satisfy*

$$\zeta^{\pm}(x) = [Pr]_x,$$

where P is the spectral projector of M for the eigenvalue 1*, that is P is the unique matrix such that*

$$P = P^2, \quad \operatorname{Im} P = \ker(I - M), \quad \ker P = \operatorname{Im}(I - M), \quad \textit{and} \quad P = PM = MP \ .$$

Proof. Same proof as for Proposition 5.17 for the time average criteria. □

Since $\zeta^{\pm}(x) = \lim_{\alpha\to 1^-}(1-\alpha)(I-\alpha M)^{-1}$, which is related to the resolvent matrix of M, one can even obtain the following more precise result, which implies that the map $\alpha \mapsto (1-\alpha)(I-\alpha M)^{-1}$ is analytic around $\alpha = 1$, a property which will be used later.

Theorem 5.32. *Given a Markov matrix M over the state space $\mathcal{E}$. We have the following asymptotic expansion around $\alpha = 1$:*

$$(I - \alpha M)^{-1} = \frac{1}{1-\alpha}P - \sum_{k=0}^{\infty}(1-\alpha)^k(\alpha^{-1}S)^{k+1} \ , \tag{5.21}$$

where P is the spectral projector of M for the eigenvalue 1*, that is P is the unique matrix such that*

$$P = P^2, \quad \operatorname{Im} P = \ker(I - M), \quad \ker P = \operatorname{Im}(I - M), \quad \textit{and} \quad P = PM = MP \ ,$$

and S satisfies $S(M-I) = (M-I)S = I - P$.

Proof. See [EC3]. □

Let us consider now the controlled case. Denote by v_α the value of the discounted infinite horizon problem. From dynamic programming equation for MDP with infinite horizon discounted criteria, Theorem 4.5, v_α is the unique solution of the stationary equation $v_\alpha = \mathcal{B}(\alpha v_\alpha)$.

Proposition 5.33. *Under* (A5)*, we have*

1. $\lim_{\alpha\to 1^-}(1-\alpha)v_\alpha = \rho\mathbf{1}$.
2. $v_\alpha - \frac{\rho}{1-\alpha}\mathbf{1}$ *is bounded (w.r.t. $\alpha \in [0,1)$).*
3. *For any converging subsequence of $v_\alpha - \frac{\rho}{1-\alpha}\mathbf{1}$ when α goes to* 1*, the limit v satisfies the ergodic Bellman equation* (5.17)*.*

Proof. **Proof of Point 2.** Denote $w_\alpha = v_\alpha - \frac{\rho}{1-\alpha}\mathbf{1}$. Since $v_\alpha = \mathcal{B}(\alpha v_\alpha)$, we obtain, using the additive homogeneity of $\mathcal{B}$:

$$\begin{aligned} w_\alpha &= \mathcal{B}(\alpha v_\alpha) - \frac{\rho}{1-\alpha}\mathbf{1} \\ &= \mathcal{B}(\alpha w_\alpha + \alpha\frac{\rho}{1-\alpha}\mathbf{1}) - \frac{\rho}{1-\alpha}\mathbf{1} \\ &= \mathcal{B}(\alpha w_\alpha) + \alpha\frac{\rho}{1-\alpha}\mathbf{1} - \frac{\rho}{1-\alpha}\mathbf{1} \\ &= \mathcal{B}(\alpha w_\alpha) - \rho\mathbf{1} \ . \end{aligned}$$

Substracting ergodic equation $v = \mathcal{B}(v) - \rho\mathbf{1}$, we obtain

$$w_\alpha - v = \mathcal{B}(\alpha w_\alpha) - \mathcal{B}(v) \ .$$

Using the nonexpansivity (1-Lipschitz continuity) of $\mathcal{B}$, we deduce:

$$\begin{aligned}\|w_\alpha - v\|_\infty &= \|\mathcal{B}(\alpha w_\alpha) - \mathcal{B}(v)\|_\infty \\ &\leq \|\alpha w_\alpha - v\|_\infty \\ &\leq \|\alpha w_\alpha - \alpha v\|_\infty + \|\alpha v - v\|_\infty \\ &\leq \alpha\|w_\alpha - v\|_\infty + (1-\alpha)\|v\|_\infty \ .\end{aligned}$$

Then,

$$\|w_\alpha - v\|_\infty \leq \|v\|_\infty$$

and so

$$\|w_\alpha\|_\infty \leq 2\|v\|_\infty \ ,$$

which shows Point 2.

Point 2. implies Point 1.

Proof of Point 3. Since closed bounded sets of $\mathbb{R}^{\mathcal{E}}$ are compact, any sequence w_{α_n} with $\alpha_n \to 1^-$ has a converging subsequence. Let us denote also by w_{α_n} such a converging sequence, and let w be its limit. We have

$$w_{\alpha_n} = \mathcal{B}(\alpha_n w_{\alpha_n}) - \rho\mathbf{1} \ .$$

Passing to the limit when n goes to ∞ in this equation, and using the continuity of $\mathcal{B}$, we obtain

$$w = \mathcal{B}(w) - \rho\mathbf{1} \ ,$$

that is w is solution of the ergodic equation. □

In the uncontrolled case, the ergodicity of the Markov chain, which is equivalent to the property that its matrix has a unique final class, implied a stronger property than (A5), which can be generalized as follows in the controlled case:

(A6) There exists $\rho \in \mathbb{R}$ and $v \in \mathbb{R}^{\mathcal{E}}$ satisfying the ergodic Bellman equation (5.17). Moreover, $\rho \in \mathbb{R}$ is unique and v is unique up to an additive constant, meaning that if v, v' satisfy (5.17), then $v - v'$ is a constant vector.

Corollary 5.34. *If* (A6) *holds, then* $v_\alpha - v_\alpha(x_0)\mathbf{1}$ *has a limit* v *when* $\alpha \to 1^-$ *and* v *satisfies (with* ρ*) the ergodic Bellman equation* (5.17)*.*

Definition 5.35. A solution v of the ergodic Bellman equation (5.17) is called a *bias* or a *relative value function* of the MDP with mean-payoff criteria.

Proof of Corollary 5.34. By Proposition 5.33, $w_\alpha = v_\alpha - \frac{\rho}{1-\alpha}\mathbf{1}$ is bounded w.r.t. $\alpha \in [0,1)$, and any limit point of w_α when $\alpha \to 1^-$ is solution of the ergodic Bellman equation (5.17). Hence, the difference between two values of the vector w_α is also bounded, so $v_\alpha - v_\alpha(x_0)\mathbf{1} = w_\alpha - w_\alpha(x_0)\mathbf{1}$ is bounded. Moreover any limit point of a sequence $v_{\alpha_n} - v_{\alpha_n}(x_0)\mathbf{1}$, with $\alpha_n \to 1^-$, is equal

to $v - v(x_0)\mathbf{1}$ for some solution v of the ergodic Bellman equation (5.17). Since $\mathcal{B}$ is additively homogeneous, $w = v - v(x_0)\mathbf{1}$ is also a solution of the ergodic Bellman equation (5.17). Moreover w satisfies $w(x_0) = 0$. By (A6), such a solution w is unique, since if w, w' satisfy (5.17) and the condition $w(x_0) = 0 = w'(x_0)$, then $w - w'$ is a constant vector and satisfies $(w - w')(x_0) = 0$, so $w = w'$. So all limit points of $v_{\alpha_n} - v_{\alpha_n}(x_0)\mathbf{1}$ are equal to this unique solution w. This implies that $v_\alpha - v_\alpha(x_0)\mathbf{1}$ converges towards w when $\alpha \to 1^-$. □

However, Assumption (A6) is not needed in the case of finite action spaces.

Proposition 5.36. *Under* (A5)*, if the sets $\mathcal{C}(x)$ are finite, then $v_\alpha - \frac{\rho}{1-\alpha}\mathbf{1}$ has a limit v when $\alpha \to 1^-$ and v satisfies (with ρ) the ergodic Bellman equation* (5.17)*. Moreover, the same property holds for $v_\alpha - v_\alpha(x_0)\mathbf{1}$.*

Proof. Let $w_\alpha = v_\alpha - \frac{\rho}{1-\alpha}\mathbf{1}$, we have $w_\alpha = \mathcal{B}(\alpha w_\alpha) - \rho\mathbf{1}$. By the continuity of $\mathcal{B}$ and the uniqueness of the solution of the previous equation, the map $\alpha \in [0,1) \mapsto w_\alpha$ is continuous. Indeed, for instance for all $0 \leq \alpha, \alpha' < 1$, we have $\|w_\alpha - w_{\alpha'}\| = \|\mathcal{B}(\alpha w_\alpha) - \mathcal{B}(\alpha w_{\alpha'})\| \leq \|\alpha w_\alpha - \alpha' w_{\alpha'}\| \leq |\alpha - \alpha'|\|w_\alpha\| + \alpha'\|w_\alpha - w_{\alpha'}\|$. So $\|w_\alpha - w_{\alpha'}\| \leq \frac{|\alpha-\alpha'|}{1-\alpha'}\|w_\alpha\|$.

Since the sets $\mathcal{C}(x)$ are finite, for any $\alpha < 1$, there exist $\pi \in \Pi$ such that $\mathcal{B}(\alpha w_\alpha) = \mathcal{B}^{(\pi)}(\alpha w_\alpha)$.

Therefore, w_α is solution of $w_\alpha = r^{(\pi)} + \alpha M^{(\pi)} w_\alpha - \rho\mathbf{1}$, and is thus given by:

$$w_\alpha = w_\alpha^\pi := (I - \alpha M^{(\pi)})^{-1}(r^{(\pi)} - \rho\mathbf{1}) \ .$$

The map $\alpha \in [0,1) \mapsto w_\alpha^\pi$ is rational and thus meromorphic (see Theorem 5.32). So, for any two feedback policies π and π', and any $x \in \mathcal{E}$, the set of zeros of the map $\alpha \in [0,1) \mapsto w_\alpha^\pi(x) - w_\alpha^{\pi'}(x)$ has no accumulation point, or the map is identically zero.

Hence, there exists $\alpha_0 < 1$ such that for all $x \in \mathcal{E}$, and any two feedback policies π and π', either $w_\alpha^\pi(x) \neq w_\alpha^{\pi'}(x)$, for all $\alpha \in [\alpha_0, 1)$, or $w_\alpha^\pi(x) = w_\alpha^{\pi'}(x)$, for all $\alpha \in [\alpha_0, 1)$.

For all $\alpha \in [\alpha_0, 1)$, there exists $\pi \in \Pi$ such that $w_\alpha = w_\alpha^\pi$. Moreover, $w_\alpha(x) \geq w_\alpha^\pi(x)$ for all $\alpha \in [0,1)$, $x \in \mathcal{E}$ and $\pi \in \Pi$. Pick one π such that $w_{\alpha_0} = w_{\alpha_0}^\pi$. If there exists $\alpha_1 \in (\alpha_0, 1)$ such that $w_{\alpha_1} \neq w_{\alpha_1}^\pi$, then there exist $\pi' \in \Pi$ and $x \in \mathcal{E}$ such that $w_{\alpha_1} = w_{\alpha_1}^{\pi'}$ and $w_{\alpha_1}(x) \neq w_{\alpha_1}^\pi(x)$. Hence $w_\alpha^{\pi'}(x) \neq w_\alpha^\pi(x)$, for all $\alpha \in [\alpha_0, 1)$ by the previous property. Since $w_\alpha(x) \geq w_\alpha^\pi$ for all $\pi \in \Pi$, we deduce that $w_\alpha^{\pi'}(x) - w_\alpha^\pi(x)$ is > 0 for $\alpha = \alpha_1$ and ≤ 0 for $\alpha = \alpha_0$. By the continuity of $\alpha \mapsto w_\alpha^{\pi'}(x) - w_\alpha^\pi(x)$, this implies that there is a zero, a contradiction. Therefore, $w_\alpha = w_\alpha^\pi$ for all $\alpha \in [\alpha_0, 1)$.

The asymptotics expansion of w_α^π has the form:

$$w_\alpha^\pi = \frac{1}{1-\alpha} v_{-1} + v + \mathcal{O}(1-\alpha)$$

for some $v_{-1}, v \in \mathbb{R}^{\mathcal{E}}$. Since $w_\alpha^\pi = w_\alpha$ which is bounded, we get that the first term is zero, and so w_α converges towards v. Passing to the limit in the equation $w_\alpha = \mathcal{B}(\alpha w_\alpha) - \rho\mathbf{1}$, we get that v is solution of the ergodic equation.

Since $v_\alpha - v_\alpha(x_0)\mathbf{1} = w_\alpha - w_\alpha(x_0)\mathbf{1}$, we get that $v_\alpha - v_\alpha(x_0)\mathbf{1}$ converges towards $v - v(x_0)\mathbf{1}$ which is also solution of the ergodic equation. □

5.3.4 An existence result

Definition 5.37 (Blackwell 62)**.** We say that a policy π is *Blackwell optimal* for the MDP, if there exists $\alpha_0 < 1$ such that for all $\alpha \in [\alpha_0, 1)$, π is optimal for the MDP with discounted infinite horizon criteria with discount factor α, that is if $v_\alpha = \mathcal{B}(\alpha v_\alpha)$ then $v_\alpha = \mathcal{B}^{(\pi)}(\alpha v_\alpha)$.

In the proof of Proposition 5.36, we have indeed shown the following result.

Theorem 5.38. *If the sets $\mathcal{C}(x)$ are finite, then there exists a Blackwell optimal policy for the MDP.* □

Definition 5.39. Consider a MDP with state space $\mathcal{E}$, action spaces $\mathcal{C}(x)$ and transition probabilities M^u_{xy}. We define the *digraph of the MDP* as the set of nodes $\mathcal{E}$ with an arc from $x \in \mathcal{E}$ to $y \in \mathcal{E}$ if there exists $u \in \mathcal{C}(x)$ such that $M^u_{xy} > 0$.

Theorem 5.40 (Bather, 73)**.** *Assume the graph of the MDP is strongly connected. Then, for all reward functions r, there exists a solution (ρ, v) to the ergodic Bellman equation* (5.17)*.*

One proof relies on metric techniques: a solution is a fixed point to an operator on a projective space. We shall give another one which relies on Blackwell strategies and works when the sets $\mathcal{C}(x)$ are finite.

Proof of Bather theorem, Theorem 5.40. Assume that Blackwell strategies exist. Let $\alpha_0 \in [0, 1)$ and $\pi \in \Pi$ be such that for all $\alpha \in [\alpha_0, 1)$, the value function v_α solution of $v_\alpha = \mathcal{B}(\alpha v_\alpha)$ satisfies $v_\alpha = \mathcal{B}^{(\pi)}(\alpha v_\alpha)$. Therefore, $v_\alpha = r^{(\pi)} + \alpha M^{(\pi)} v_\alpha$, and is thus given by:

$$v_\alpha = v^\pi_\alpha := (I - \alpha M^{(\pi)})^{-1} r^{(\pi)} \ .$$

By Theorem 5.32, we have

$$v_\alpha = \frac{1}{1-\alpha} v_{-1} + v_0 + \mathcal{O}(1-\alpha) \ ,$$

for some vectors $v_{-1}, v_0 \in \mathbb{R}^{\mathcal{E}}$.

Since the graph of the MDP is strongly connected, one can construct a Markov strategy (that is a relaxed policy) with an irreducible associated transition matrix. Indeed, for any arc (x, y) of the graph of the MDP, there exist $u \in \mathcal{C}(x)$ such that $M^{(u)}_{xy} > 0$. Denote by G_x a finite subset of $\mathcal{C}(x)$ composed of the actions u associated to the arcs (x, y) of the MDP. Considering a policy $\tilde{\pi} \in \Pi^{\mathrm{R}}$ such that $\tilde{\pi}(x)$ is the uniform probability on G_x, we get that the transition matrix of the Markov chain X_n induced by this policy satisfies $M^{(\tilde{\pi})}_{xy} = \frac{1}{\mathrm{card}(G_x)} \sum_{u \in G_x} M^{(u)}_{xy} > 0$ for all arcs (x, y) of the MDP. Then, the graph of $M^{(\tilde{\pi})}$ coincides with the graph of the MDP, and thus $M^{(\tilde{\pi})}$ is irreducible.

We have $v_\alpha = \mathcal{B}(\alpha v_\alpha) \geq \mathcal{B}^{(\tilde{\pi})}(\alpha v_\alpha)$. Using the asymptotic expansion of v_α, we get:

$$\frac{1}{1-\alpha} v_{-1} + v_0 + \mathcal{O}(1-\alpha) \geq r^{(\tilde{\pi})} + \alpha M^{(\tilde{\pi})} \left(\frac{1}{1-\alpha} v_{-1} + v_0 + \mathcal{O}(1-\alpha) \right) \ .$$

Taking the dominant terms, we obtain:

$$v_{-1} \geq M^{(\tilde{\pi})} v_{-1}$$

which by Perron-Frobenius theorem applied to the irreducible Markov matrix $M^{(\tilde{\pi})}$ implies that v_{-1} is of the form $\rho\mathbf{1}$ for some $\rho \in \mathbb{R}$ (since $\mathbf{1}$ is the Perron vector of $M^{(\tilde{\pi})}$). Then, the asymptotic expansion of v_α has the form:

$$v_\alpha = \frac{\rho}{1-\alpha}\mathbf{1} + v_0 + \mathcal{O}(1-\alpha) \ .$$

Puting this in the equation $v_\alpha = \mathcal{B}(\alpha v_\alpha)$, and using that $\mathcal{B}$ is additively homogeneous, we get

$$\rho\mathbf{1} + v_0 + \mathcal{O}(1-\alpha) = \mathcal{B}(\alpha v_0 + \mathcal{O}(1-\alpha)) \ .$$

Since $\mathcal{B}$ is Lipschitz continuous, we deduce that $\rho\mathbf{1} + v_0 = \mathcal{B}(v_0) + \mathcal{O}(1-\alpha)$ and so $\rho\mathbf{1} + v_0 = \mathcal{B}(v_0)$, so v_0 is solution of the ergodic equation. □

5.3.5 Policy iteration algorithm

Recall that the value v_α of the discounted infinite horizon problem is the unique solution of the stationary dynamic programming equation $v_\alpha = \mathcal{B}(\alpha v_\alpha)$ and that the policy iteration (or Howard) algorithm to solve this equation consists in the following successive steps $k \geq 0$, starting from a policy $\pi^0 \in \Pi$ (see Definition 4.11):

1. w^k is the unique solution of the Kolmogorov equation associated to the policy π^k:

$$v(x) = r(x, \pi^k(x)) + \alpha \sum_{y \in \mathcal{E}} M_{xy}^{(\pi^k(x))} v(y) \quad \forall x \in \mathcal{E} \ .$$

2. π^{k+1} is an optimal policy for w^k, that is an element π such that

$$\pi(x) \in \underset{u \in \mathcal{C}(x)}{\text{Argmax}} \left(r(x,u) + \alpha \sum_{y \in \mathcal{E}} M_{xy}^{(u)} w^k(y) \right) \quad \forall x \in \mathcal{E} \ .$$

which can also be rewritten in functional form :

1. w^k is the solution of the equation $w = \mathcal{B}^{(\pi^k)}(\alpha w)$.
2. π^{k+1} is an element π of Π such that $\mathcal{B}^{(\pi)}(\alpha w^k) = \mathcal{B}(\alpha w^k)$.

One may thus ask what is the limit of the algorithm when $\alpha \to 1^-$. Let us first study the behavior of the algorithm under the following assumption.

(A7) Assume that all the matrices $M^{(\pi)}$, with $\pi \in \Pi$, are ergodic (have a unique final class).

In that case, we have

- for all $\pi \in \Pi$, the ergodic equation

$$\rho\mathbf{1} + v = \mathcal{B}^{(\pi)}(v)$$

has a solution (ρ^π, v^π) with $\rho^\pi \in \mathbb{R}$ and $v^\pi \in \mathbb{R}^{\mathcal{E}}$. Moreover, one may force v^π to be unique by imposing $v^\pi(x_0) = 0$, for some fixed state $x_0 \in \mathcal{E}$.

- Then the solution w_α^π of $w = \mathcal{B}^{(\pi)}(\alpha w)$ satisfies:

$$w_\alpha^\pi = \frac{1}{1-\alpha}\rho^\pi \mathbf{1} + \mu \mathbf{1} + v^\pi + \mathcal{O}(1-\alpha) \ ,$$

  for some constant $\mu \in \mathbb{R}$.

- With the same arguments as in the proof of Proposition 5.36, for all π the exists π' such that $\mathcal{B}(\alpha w_\alpha^\pi) = \mathcal{B}^{(\pi')}(\alpha w_\alpha^\pi)$ for all α close to 1.

- Therefore for α close to 1, the sequence π^k of PI algorithm is independent of α, and we have

$$w^k = \frac{1}{1-\alpha}\rho^k \mathbf{1} + \mu^k \mathbf{1} + v^k + \mathcal{O}(1-\alpha) \ ,$$

  where

$$\rho^k \mathbf{1} + v^k = \mathcal{B}^{(\pi^k)}(v^k), \quad v^k(x_0) = 0 \ .$$

- Since $\mathcal{B}$ and $\mathcal{B}^{(\pi)}$ are additively homogeneous, we have: π^{k+1} is optimal for v^k in $\mathcal{B}(v^k)$.

Let us consider first the stronger assumption:

(A8) Assume that all the matrices $M^{(\pi)}$, with $\pi \in \Pi$, are irreducible.

Definition 5.41 (Policy Iteration alogirithm (PI) for irreducible matrices)**.** Assume (A8) holds, and that the optimization problems in Bellman equations can be solved, that is, for all $v \in \mathbb{R}^{\mathcal{E}}$, there exists $\pi \in \Pi$ such that $\mathcal{B}^{(\pi)}(v) = \mathcal{B}(v)$. The *policy iteration algorithm* applied to the ergodic Bellman equation $\rho\mathbf{1} + v = \mathcal{B}(v)$ consists in the following successive steps $k \geq 0$, starting from a policy $\pi^0 \in \Pi$:

1. $(\rho^k, v^k) \in \mathbb{R} \times \mathbb{R}^{\mathcal{E}}$ is the unique solution of the ergodic Kolmogorov equation associated to the policy π^k:

$$\rho + v(x) = r(x, \pi^k(x)) + \sum_{y \in \mathcal{E}} M_{xy}^{(\pi^k(x))} v(y) \quad \forall x \in \mathcal{E} \ ,$$

   satistying in addition the condition $v(x_0) = 0$ for some fixed point $x_0 \in \mathcal{E}$.

2. π^{k+1} is an optimal policy for v^k, that is an element π such that

$$\pi(x) \in \operatorname*{Argmax}_{u \in \mathcal{C}(x)} \left(r(x,u) + \sum_{y \in \mathcal{E}} M_{xy}^{(u)} v^k(y) \right) \quad \forall x \in \mathcal{E} \ .$$

Fact 5.42. The *policy iteration algorithm* applied to the ergodic Bellman equation $\rho\mathbf{1} + v = \mathcal{B}(v)$ consists in the following successive steps, starting from a policy $\pi^0 \in \Pi$: for $k \geq 0$, do

1. ρ^k is the value and v^k is the biais of problem when the policy is frozen to π^k that is the solution of the equation $\rho\mathbf{1} + v = \mathcal{B}^{(\pi^k)}(v)$ (with $v(x_0) = 0$).

2. π^{k+1} is an optimal policy for v^k, that is an element π of Π such that $\mathcal{B}^{(\pi)}(v^k) = \mathcal{B}(v^k)$.

Remark 5.43. Recall that for the PI for discounted problems, we proved that the sequence of values satisfies

$$w^k \leq w^{k+1} \leq \cdots \leq v \ ,$$

and

$$\lim_{k\to\infty} w^k = v \ .$$

Since $w^k = \frac{1}{1-\alpha}\rho^k \mathbf{1} + \mu^k \mathbf{1} + v^k + \mathcal{O}(1-\alpha)$, we obtain:

$$\rho^k \leq \rho^{k+1} \cdots \leq \rho \ ,$$

and if $\rho^k = \rho^{k+1}$ then

$$\mu^k \mathbf{1} + v^k \leq \mu^{k+1}\mathbf{1} + v^{k+1} \ .$$

Theorem 5.44. *Assume* (A8) *holds, and that the optimization problems in Bellman equations can be solved, that is, for all $v \in \mathbb{R}^{\mathcal{E}}$, $\mathcal{B}(v)$ is finite and there exists $\pi \in \Pi$ such that $\mathcal{B}^{(\pi)}(v) = \mathcal{B}(v)$. Let ρ^k, v^k, π^k be the sequence generated by PI algorithm. We have, for all $k \geq 0$,*

$$\rho^k \leq \rho^{k+1} \ .$$

Moreover, there exists a solution $(\rho, v) \in \mathbb{R} \times \mathbb{R}^{\mathcal{E}}$ to the ergodic equation: $\rho\mathbf{1} + v = \mathcal{B}(v)$, and we have, for all $k \geq 0$,

$$\rho^k \leq \rho^{k+1} \leq \cdots \leq \rho \ .$$

Proposition 5.45 (Sub or supersolutions)**.** *Let $\mathcal{B}$ be a monotone additively homogeneous operator from $\mathbb{R}^{\mathcal{E}}$ to itself, beeing the Bellman operator of a undiscounted MDP. Assume that there exists $(\rho, v) \in \mathbb{R} \times \mathbb{R}^{\mathcal{E}}$ solution of the ergodic equation $\rho\mathbf{1} + v = \mathcal{B}(v)$. Then, for $(\zeta, w) \in \mathbb{R} \times \mathbb{R}^{\mathcal{E}}$, we have*

$$\zeta\mathbf{1} + w \leq \mathcal{B}(w) \implies \zeta \leq \rho \tag{5.22}$$

$$\zeta\mathbf{1} + w \geq \mathcal{B}(w) \implies \zeta \geq \rho \ . \tag{5.23}$$

Proof. From $\zeta\mathbf{1}+w \leq \mathcal{B}(w)$ and the monotonicity and additive homogeneity, we get that $T\zeta\mathbf{1}+w \leq \mathcal{B}^T(w)$ for all $t \geq 1$.

Since $\mathcal{B}^T(w)$ is the value function of the finite horizon problem, we obtain (from Proposition 5.27) that $\lim_{T\to\infty} \frac{1}{T}\mathcal{B}^T(w) = \rho$.

Passing to the limit into the previous inequality, we deduce that $\zeta \leq \rho$.

The same holds for the reverse inequality. □

Proof of Theorem 5.44. Since all the $M^{(\pi)}$ are irreducible, the associated ergodic equations have a solution. We have

$$\rho^k\mathbf{1} + v^k = \mathcal{B}^{(\pi^k)}(v^k) \tag{5.24}$$

$$\mathcal{B}(v^k) = \mathcal{B}^{(\pi^{k+1})}(v^k) \tag{5.25}$$

$$\rho^{k+1}\mathbf{1} + v^{k+1} = \mathcal{B}^{(\pi^{k+1})}(v^{k+1}) \ . \tag{5.26}$$

Using the first and second equations together with the definition of $\mathcal{B}$, we get

$$\rho^k\mathbf{1} + v^k = \mathcal{B}^{(\pi^k)}(v^k) \leq \mathcal{B}(v^k) = \mathcal{B}^{(\pi^{k+1})}(v^k) \ .$$

Hence (ρ^k, v^k) is a subsolution of the ergodic Kolmogorov equation $\rho\mathbf{1} + v = \mathcal{B}^{(\pi^{k+1})}(v)$. From (5.26), (ρ^{k+1}, v^{k+1}) is a solution of this ergodic equation.

From (5.22), we deduce that $\rho^k \leq \rho^{k+1}$.

We also have $\rho^k\mathbf{1} + v^k \leq \mathcal{B}(v^k)$, so (ρ^k, v^k) is a subsolution of the ergodic Bellman equation.

Since the graph of the MDP contains the one of any matrix $M^{9\pi)}$, it is strongly connected. Then, by Bather theorem, there exists a solution $(\rho, v) \in \mathbb{R} \times \mathbb{R}^{\mathcal{E}}$ to the ergodic equation: $\rho\mathbf{1} + v = \mathcal{B}(v)$. Moreover, applying (5.22) to $\mathcal{B}$, we deduce that $\rho^k \leq \rho$, for all $k \geq 0$. □

Theorem 5.46. *If in the PI algorithm for irreducible matrices, we have $\rho^k = \rho^{k+1}$, then*

$$v^k = v^{k+1} \ .$$

Proof. Assume that in the PI algorithm, we have $\rho^k = \rho^{k+1}$. Recall that we proved during the proof of the first properties:

$$\rho^k\mathbf{1} + v^k = \mathcal{B}^{(\pi^k)}(v^k) \leq \mathcal{B}(v^k) = \mathcal{B}^{(\pi^{k+1})}(v^k) \ .$$

Since $\rho^{k+1}\mathbf{1} + v^{k+1} = \mathcal{B}^{(\pi^{k+1})}(v^{k+1})$, taking the difference, we obtain

$$v^k - v^{k+1} \leq M^{(\pi^{k+1})}(v^k - v^{k+1}) \ .$$

By Perron-Frobenius theorem applied to the irreducible matrix $M^{(\pi^{k+1})}$, we obtain that $v^k - v^{k+1}$ is a constant vector.

Since in addition $(v^k - v^{k+1})(x_0) = 0$, we obtain $v^k = v^{k+1}$.

□

Corollary 5.47. *If the sets $\mathcal{C}(x)$ are finite, then the PI algorithm for irreducible matrices converges after a finite number of iterations.*

Proof. If the sets $\mathcal{C}(x)$ are finite, the set of feedback policies, Π is finite.

Therefore, there exists $k < \ell$ such that $\pi^k = \pi^\ell$.

From the uniqueness of the solution (ρ, v) to the equation $\mathcal{B}^{(\pi^k)}(w) = w + \rho\mathbf{1}$, with the additional condition $w(x_0) = 0$, we get that $\rho^k = \rho^\ell$ and $v^k = v^\ell$.

Since the sequence ρ^k is nondecreasing, $\rho^k \leq \rho^{k+1} \leq \cdots \leq \rho^\ell$, and satisfies $\rho^k = \rho^\ell$, we get the equality $\rho^k = \rho^{k+1} = \cdots = \rho^\ell$.

This implies $v^k = v^{k+1}$ by Theorem 5.46.

Since π^{k+1} is optimal for $v^k = v^{k+1}$, and (ρ^{k+1}, v^{k+1}) is a solution of the ergodic equation associated to π^{k+1}, we obtain

$$\rho^{k+1}\mathbf{1} + v^{k+1} = \mathcal{B}^{(\pi^{k+1})}(v^{k+1}) = \mathcal{B}^{(\pi^{k+1})}(v^k) = \mathcal{B}(v^k) = \mathcal{B}(v^{k+1})$$

and so (ρ^{k+1}, v^{k+1}) is a solution to the ergodic Bellman equation.

Hence $\rho^{k+1} = \rho$ and π^{k+1} is optimal. □

Definition 5.48 (Policy Iteration algorithm (PI) for ergodic matrices)**.** Assume (A7) holds. The *policy iteration algorithm* applied to the ergodic Bellman equation $\rho\mathbf{1} + v = \mathcal{B}(v)$ consists in the following successive steps $k \geq 0$, starting from a policy $\pi^0 \in \Pi$:

1. $(\rho^k, v^k) \in \mathbb{R} \times \mathbb{R}^{\mathcal{E}}$ is the unique solution of the ergodic Kolmogorov equation associated to the policy π^k:
$$\rho + v(x) = r(x, \pi^k(x)) + \sum_{y \in \mathcal{E}} M_{xy}^{(\pi^k(x))} v(y) \quad \forall x \in \mathcal{E} \ ,$$
satistying in addition the condition
$$mv = 0 \textit{where } m \textit{ is the unique invariant probability measure of } M^{(\pi^k)}.$$

2. π^{k+1} is an optimal policy for v^k, that is an element π such that
$$\pi(x) \in \operatorname*{Argmax}_{u \in \mathcal{C}(x)} \left(r(x,u) + \sum_{y \in \mathcal{E}} M_{xy}^{(u)} v^k(y) \right) \quad \forall x \in \mathcal{E} \ ,$$
such that $\pi(x) = \pi^k(x)$ whenever possible (conservative policy improvement).

Theorem 5.49. *Assume* (A7) *holds, and that the optimization problems in Bellman equations can be solved, that is, for all $v \in \mathbb{R}^{\mathcal{E}}$, $\mathcal{B}(v)$ is finite and there exists $\pi \in \Pi$ such that $\mathcal{B}^{(\pi)}(v) = \mathcal{B}(v)$. Let ρ^k, v^k, π^k be the sequence generated by PI algorithm. We have, for all $k \geq 0$,*
$$\rho^k \leq \rho^{k+1} \ .$$
Moreover, if there exists a solution $(\rho, v) \in \mathbb{R} \times \mathbb{R}^{\mathcal{E}}$ to the ergodic equation: $\rho \mathbf{1} + v = \mathcal{B}(v)$, we also have, for all $k \geq 0$,
$$\rho^k \leq \rho^{k+1} \leq \cdots \leq \rho \ .$$

Theorem 5.50. *If in the PI algorithm for ergodic matrices, we have $\rho^k = \rho^{k+1}$, then*
$$v^k \leq v^{k+1} \ .$$

Corollary 5.51. *If the sets $\mathcal{C}(x)$ are finite, then the PI algorithm for ergodic matrices converges after a finite number of iterations.*

Proof. If the sets $\mathcal{C}(x)$ are finite, the set of feedback policies, Π is finite.

Therefore, there exists $k < \ell$ such that $\pi^k = \pi^\ell$.

From the uniqueness of the solution (ρ, v) to the equation $\mathcal{B}^{(\pi^k)}(w) = w + \rho \mathbf{1}$, with the additional condition $mw = 0$, we get that $\rho^k = \rho^\ell$ and $v^k = v^\ell$.

Since the sequence ρ^k is nondecreasing, $\rho^k \leq \rho^{k+1} \leq \cdots \leq \rho^\ell$, and satisfies $\rho^k = \rho^\ell$, we get the equality $\rho^k = \rho^{k+1} = \cdots = \rho^\ell$.

This implies $v^k \leq v^{k+1} \leq \cdots v^\ell$ by Theorem 5.50 and so the equality $v^k = v^{k+1} = \cdots = v^\ell$.

Since π^{k+1} is optimal for $v^k = v^{k+1}$, and (ρ^{k+1}, v^{k+1}) is a solution of the ergodic equation associated to π^{k+1}, we obtain
$$\rho^{k+1}\mathbf{1} + v^{k+1} = \mathcal{B}^{(\pi^{k+1})}(v^{k+1}) = \mathcal{B}^{(\pi^{k+1})}(v^k) = \mathcal{B}(v^k) = \mathcal{B}(v^{k+1})$$
and so (ρ^{k+1}, v^{k+1}) is a solution to the ergodic Bellman equation and π^{k+1} is optimal for v^{k+1}. Since the improvement is conservative, we get that $\pi^{k+2} = \pi^{k+1}$. □

The proof of Theorem 5.50 is based on the following lemma.

Lemma 5.52. *Let M be an ergodic Markov matrix with unique final class denoted F. Assume $v \leq Mv$. Then, we have:*

1. *$v = Mv$ and $v = \lambda \mathbf{1}$ on F, for some real λ, that is $v(x) = [Mv](x)$ and $v(x) = \lambda$ for all $x \in F$.*
2. *If $v = 0$ on F, that is $v(x) = 0$ for all $x \in F$, then $v \leq 0$.*

Proof. Denote v_A and M_{AB} the restriction of v to the set $A \subset \mathcal{E}$ and of M to the rows in A and columns in B respectively.

Proof of Point 1. Since $M_{FF^c} = 0$, $v \leq Mv$ implies $v_F \leq M_{FF} v_F$. Applying Perron-Frobenius theorem to the irreducible Markov matrix M_{FF}, we deduce $v_F = M_{FF} v_F$ and $v_F = \lambda \mathbf{1}_F$ for some scalar $\lambda \in \mathbb{R}$. So $v_F = M_{FF^c} v_{F^c} + M_{FF} v_F = [Mv]_F$.

Proof of Point 2. If $v_F = 0$, then $v_{F^c} \leq M_{F^c F^c} v_{F^c}$. By the nonnegativity of $M_{F^c F^c}$, we deduce that $v_{F^c} \leq (M_{F^c F^c})^n v_{F^c}$, for all $n \geq 1$.

Since F is the unique final class of M, then F^c contains only transient states. So $\rho(M_{F^c F^c}) < 1$ and $\lim_{n\to\infty} (M_{F^c F^c})^n = 0$.

Passing to the limit in $v_{F^c} \leq (M_{F^c F^c})^n v_{F^c}$, we get $v_{F^c} \leq 0$, and since $v_F = 0$, we have $v \leq 0$. □

Proof of Theorem 5.50. Assume that in the PI algorithm, we have $\rho^k = \rho^{k+1}$. Recall that we proved during the proof of the first properties:

$$\rho^k \mathbf{1} + v^k = \mathcal{B}^{(\pi^k)}(v^k) \leq \mathcal{B}(v^k) = \mathcal{B}^{(\pi^{k+1})}(v^k) \ .$$

Since $\rho^{k+1}\mathbf{1} + v^{k+1} = \mathcal{B}^{(\pi^{k+1})}(v^{k+1})$, taking the difference, we obtain

$$v^k - v^{k+1} \leq M^{(\pi^{k+1})}(v^k - v^{k+1}) \ .$$

Let F be the unique final class of the ergodic matrix $M^{(\pi^{k+1})}$. By Point 1 of Lemma 5.52, we obtain $[v^k - v^{k+1}](x) = [M^{(\pi^{k+1})}(v^k - v^{k+1})](x)$ and $(v^k - v^{k+1})(x) = \lambda$ for all $x \in F$, for some $\lambda \in \mathbb{R}$.

This implies in particular that on F, we have $\mathcal{B}(v^k) = \mathcal{B}^{(\pi^{k+1})}(v^k) = \mathcal{B}^{(\pi^{k+1})}(v^{k+1}) + \lambda \mathbf{1}_F = \rho^{k+1}\mathbf{1} + v^{k+1} + \lambda \mathbf{1}_F = \rho^k \mathbf{1} + v^k = \mathcal{B}^{(\pi^k)}(v^k)$.

So, for all $x \in F$, $\pi^k(x)$ was already optimal for v^k, and since policy improvement is conservative, we have $\pi^k(x) = \pi^{k+1}(x)$.

Then, the restriction to rows in F of $M^{(\pi^k)}$ and $M^{(\pi^{k+1})}$ are the same and since F is a final class of $M^{(\pi^{k+1})}$, it is also a final class of $M^{(\pi^k)}$.

Since the invariant measure of an ergodic matrix has a support equal to the final class, we get that the invariant measures of $M^{(\pi^k)}$ and $M^{(\pi^{k+1})}$ coincide. Let us denote it by m.

Then, the constraint "$mv = 0$" implies that $mv^k = 0$ and $mv^{k+1} = 0$. Since $(v^k - v^{k+1})(x) = \lambda$ for all $x \in F$, we deduce that $v^k - v^{k+1} = 0$ on F.

By Point 2 of Lemma 5.52, we obtain that $v^k - v^{k+1} \leq 0$. □

Aplying PI algorithm, we deduce:

Corollary 5.53. *If* (A7) *holds and the sets $\mathcal{C}(x)$ are finite, then there exists a solution to the ergodic Bellman equation.*

This includes the case of Kolmogorov equations, but also some cases in which the graph of the MDP is not strongly connected.

Corollary 5.54. *The same holds if the sets $\mathcal{C}(x)$ are finite, and if instead of assuming* (A7)*, we assume that there exists a sequence of the PI algorithm for ergodic matrices which only meet ergodic matrices.*

Example 5.55 (A simple non strongly connected example)**.** Consider

$$\begin{aligned}
\rho + v(1) &= \max(1 + \frac{1}{2}v(1) + \frac{1}{2}v(2), v(2)) \\
\rho + v(2) &= 2 + v(3) \\
\rho + v(3) &= \max(v(2), v(3))
\end{aligned}$$

This is the ergodic dynamic programming equation of a MDP with 3 states, $\mathcal{E} = \{1, 2, 3\}$, a maximum of 2 actions: $\mathcal{C}(1) = \mathcal{C}(3) = \{1, 2\}$, $\mathcal{C}(2) = \{1\}$, and the following transition probabilities:

$$\begin{aligned}
M_{1.}^{(1)} &= \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 \end{bmatrix} \\
M_{1.}^{(2)} &= \begin{bmatrix} 0 & 1 & 0 \end{bmatrix} \\
M_{2.}^{(1)} &= \begin{bmatrix} 0 & 0 & 1 \end{bmatrix} \\
M_{3.}^{(1)} &= \begin{bmatrix} 0 & 1 & 0 \end{bmatrix} \\
M_{3.}^{(2)} &= \begin{bmatrix} 0 & 0 & 1 \end{bmatrix}
\end{aligned}$$

Example 5.56 (An example in which conservative improvement is essential)**.**

$$\begin{aligned}
\rho + v(1) &= v(2) + 1 \\
\rho + v(2) &= v(3) - 1 \\
\rho + v(3) &= \max(v(1), v(3))
\end{aligned}$$

This is the ergodic dynamic programming equation of a MDP with 3 states, $\mathcal{E} = \{1, 2, 3\}$, a maximum of 2 actions: $\mathcal{C}(1) = \mathcal{C}(2) = \{1\}$, $\mathcal{C}(3) = \{1, 2\} = \mathcal{C}$, and the following transition probabilities:

$$\begin{aligned}
M_{1.}^{(1)} &= \begin{bmatrix} 0 & 1 & 0 \end{bmatrix} \\
M_{2.}^{(1)} &= \begin{bmatrix} 0 & 0 & 1 \end{bmatrix} \\
M_{3.}^{(1)} &= \begin{bmatrix} 1 & 0 & 0 \end{bmatrix} \\
M_{3.}^{(2)} &= \begin{bmatrix} 0 & 0 & 1 \end{bmatrix}
\end{aligned}$$

So there exist only 2 policies π_1 and π_2. Indeed, these are applications from $\mathcal{E}$ to $\mathcal{C}$ such that $\pi_i(1) = \pi_i(2) = 1$, for $i = 1, 2$, and $\pi_1(3) = 1$ and $\pi_2(3) = 2$. They thus satisfy:

$$M^{(\pi_1)} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}, \quad M^{(\pi_2)} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix}, \quad r^{(\pi_1)} = r^{(\pi_1)} = \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix} .$$

Their invariant measures are

$$m^{(\pi_1)} = \begin{bmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{bmatrix}, \quad m^{(\pi_2)} = \begin{bmatrix} 0 & 0 & 1 \end{bmatrix} \ .$$

Consider policy iterations in which the condition $mv = 0$ is applied in value computation step, but policy improvement may not be conservative in policy step. Starting from $\pi^0 = \pi_1$, we get necessarily $\rho^0 = m^{(\pi^0)} r^{(\pi^0)} = 0$ and thus v^0 must satisfy $v(1) = v(2) + 1$, $v(2) = v(3) - 1$, $v(3) = v(1)$, and $m^{(\pi_1)} v = (v(1) + v(2) + v(3))/3 = 0$, which leads to $v^0 = \begin{bmatrix} 1/3 & -2/3 & 1/3 \end{bmatrix}^T$. Then, π_1 and π_2 are both optimal. If we choose π^1 in a conservative way, then one must choose $\pi^1 = \pi_1$, and the algorithm stops, since $\pi^1 = \pi^0$, $\rho^1 = \rho^0$ and $v^1 = v^0$. However, if we choose $\pi^1 = \pi_2$, then $\rho^1 = \rho^0 = 0$, but v^1 must satisfy $v(1) = v(2) + 1$, $v(2) = v(3) - 1$, $v(3) = v(3)$, and $m^{(\pi_2)} v = v(3) = 0$, which leads to $v^1 = \begin{bmatrix} 0 & -1 & 0 \end{bmatrix}^T = v^0 - \frac{1}{3}\mathbf{1}$, and so to the same choice of policies for π^2. If we alternate this choice at each step of the algorithm, then the algorithm never converge.

Example 5.57 (Blackmailer)**.** Consider a blackmailer who is blackmailing a victim each day, by asking her a certain amount of money U_t depending on time $t \in \mathbb{N}$. At each time $t \in \mathbb{N}$, the victim may be willing or not, but if she is not willing at some time, then she will neither be willing anymore. We shall denote by

X_t : the state of the victim at time t, where $X_t = 1$ if she is willing, and $X_t = 0$ otherwise;

U_t : the amount of money asked by the blackmailer at time t, where we assume that $U_t \leq 1$.

We then consider a MDP in which

- the state space is $\mathcal{E} = \{0, 1\}$.
- the action spaces is $\mathcal{C} = [0, 1] = \mathcal{C}(x)$ (note that $\mathcal{C}$ is infinite but compact).
- the dynamics satisfies

  $$M_{01}^{(u)} = \mathbb{P}\left(X_{t+1} = 1 \mid X_t = 0, U_t = u\right) = 0, \quad M_{00}^{(u)} = \mathbb{P}\left(X_{t+1} = 0 \mid X_t = 0, U_t = u\right) = 1 \ .$$

  We shall assume that

  $$M_{10}^{(u)} = \mathbb{P}\left(X_{t+1} = 0 \mid X_t = 1, U_t = u\right) = u^2, \quad M_{11}^{(u)} = \mathbb{P}\left(X_{t+1} = 1 \mid X_t = 1, U_t = u\right) = 1 - u^2 \ ,$$

- the reward of the blackmailer at time t is equal to

  $$r(x; u) = xu \ .$$

Consider first the discounted infinite horizon criteria with discount factor $0 < \alpha < 1$, so that the expected payoff of the blackmailer that he want to maximize is

$$\mathbb{E}\left[\sum_{t=0}^{\infty} \alpha^t r(X_t, U_t)\right] \ .$$

Let $v_\alpha(x)$ be the value function at state $x \in \mathcal{E}$ of the above MDP. Then, the Dynamic programming equation satisfied by v_α is:

$$\begin{aligned} v_\alpha(0) &= \alpha v_\alpha(0) \\ v_\alpha(1) &= \max_{u\in[0,1]} \{u + \alpha(u^2 v_\alpha(0) + (1-u^2)v_\alpha(1))\} \ . \end{aligned}$$

Hence, $v_\alpha(0) = 0$ and $w_\alpha := v_\alpha(1)$ satisfies:

$$w_\alpha = \max_{u\in[0,1]} \{u + \alpha(1-u^2)w_\alpha\} = \alpha w_\alpha + \frac{1}{4\alpha w_\alpha} \ .$$

Therefore, the optimal control among all $u > 0$ is $u_\alpha = 1/(2\alpha w_\alpha)$ which is ≤ 1 if $w_\alpha \geq 1/(2\alpha)$. In that case

$$w_\alpha = \frac{1}{2\sqrt{\alpha(1-\alpha)}} \ .$$

So these formula hold when $\alpha \geq 1/2$, and in that case an optimal stationary policy $\pi_\alpha(1)$ to be applied at each time in which the state X_t is equal to 1 is given by

$$\pi_\alpha(1) = u_\alpha = \sqrt{\frac{1-\alpha}{\alpha}}$$

Therefore $\lim_{\alpha\to 1^-} w_\alpha = +\infty$ and $\lim_{\alpha\to 1^-}((1-\alpha)w_\alpha) = 0$. This implies that the ergodic equation has no solution (otherwise, $\rho = 0$ and w_α would have been bounded). Moreover, $\lim_{\alpha\to 1^-} \pi_\alpha(1) = 0$, and so the optimal stationary policy in the long run would be to ask nothing.

Similarly, if we consider the maximization of the mean-payoff criteria:

$$\lim_{T\to\infty} \frac{1}{T}\mathbb{E}\left[\sum_{t=0}^{T-1} r(X_t; U_t) \mid X_0 = x\right] \ ,$$

then the limit of the optimal policy (at time 0) is 0.

Note that the Markov matrix associated to the policy $\pi(1) = 0$ has two final classes, so that the policy iteration cannot be applied apriori.

Exercise 5.3.1. Consider the blackmailer in which we restrict the action space $\mathcal{C}$ to be $\mathcal{C} = [\varepsilon, 1]$. Show that the matrices associated to all policies are ergodic ($\{0\}$ is the unique final class), and that policy iteration stops after a finite number of steps, although the action space is infinite.

5.4 Risk sensitive control

5.4.1 Motivation

In mathematical finance, one is interested in the optimization of the *total wealth* or rather of the *return factor or return rate* of a "portfolio", that is

$$\frac{W_T}{W_0} \quad \text{or} \quad \log\frac{W_T}{W_0} \ .$$

where W_k is the total wealth at time k.

The return in one time unit can generally be written in the form:

$$\frac{W_{k+1}}{W_k} = R(X_k, U_k, D_{k+1}) \quad \text{where} \quad X_{k+1} = f(X_k, U_k, D_{k+1}),$$

for some maps R, f, where X_k is the "state" of the portfolio, for instance the proportion in each asset (stock) or any financial product, U_k is the action of the investor, like purchasing orders, and D_k is the random disturbances in the parameters of the portfolio, like return rates.

When the D_k are either independent random variables or a Markov chain, the process (Y_k, U_k) with $Y_k = (X_k, D_k)$ can be seen as a Markov Decision Process.

It is stationary if $(D_k)_{k\geq 0}$ is stationary.

Then, one may wish to maximize the expectation of any increasing function φ of the random return rate in T time units:

$$\mathbb{E}\left[\varphi(\log \frac{W_T}{W_0})\right] \tag{5.27}$$

and to its limit when T goes to infinity.

The choice of φ will depend on the risk one wish to take. Examples are:

- $\varphi(x) = \frac{1}{\gamma}\exp(\gamma x)$, where $1 - \gamma \in \mathbb{R}$ is called the risk aversion parameter.
- When $\gamma = 0$, the optimization is the same as for $\varphi(x) = x$.
- One can replace (5.27) by considering $\rho(\log \frac{W_T}{W_0})$, where ρ is a *risk measure*, that is a real valued map on random variables, which is monotone and additively homogeneous ($\rho(X+\lambda) = \rho(X) + \lambda)$), like expectation.

We will consider the case where $\varphi(x) = \frac{1}{\gamma}\exp(\gamma x)$, with $\gamma \geq 0$ only, which reduces either to a multiplicative payoff ($\gamma > 0$) or an additive payoff ($\gamma = 0$).

5.4.2 Risk sensitive control in finite horizon

Consider a stationary Markov chain $(D_k)_{k\geq 0}$ with transition matrix M on the space $\mathcal{D}$, and a MDP on $\mathcal{E} \times \mathcal{D}$ with action spaces $\mathcal{C}(x) \subset \mathcal{C}$ and transitions given by

$$X_{k+1} = f(X_k, U_k, D_{k+1}), \text{ where} Y_k = (X_k, D_k), \ \forall k \geq 0 \ .$$

Consider a nonnegative map $R : \mathcal{E} \times \mathcal{C} \times \mathcal{D} \to \mathbb{R}_+$ and, for all strategies $\sigma = (\sigma_k)_{k\geq 0}$ in Σ or Σ^{R}, the *multiplicative payoff* with finite horizon $T \geq 1$:

$$J^{(T,\sigma)}(y) := J^T(Y;U) := \mathbb{E}\left[\left(\prod_{m=0}^{T-1} R(X_m, U_m, D_{m+1})\right)\varphi(X_T) \mid Y_0 = y\right] \ , \tag{5.28}$$

where $(Y, U) := (Y_k, U_k)_{k\geq 0}$ is the process induced by σ.

Theorem 5.58. *Assume that the map r is bounded from above. Let v^T be the value function of the Markov decision problem:*

$$v^T(y) := \max_{\sigma} J^{(T,\sigma)}(y) \ , \ \forall y \in \mathcal{E} \times \mathcal{D} \ ,$$

where the maximum is taken over all relaxed strategies starting at time 0*. Then,* v *satisfies the following forward recurrence:*

$$v^T(x,d) = \sup_{u\in\mathcal{C}(x)} \left(\sum_{z\in\mathcal{D}} M_{dz} R(x,u,z) v^{T-1}(f(x,u,z),z) \right) \quad \forall (x,d) \in \mathcal{E}\times\mathcal{D} \ . \tag{5.29}$$

with final condition $v_T(x,d) = \varphi(x)$*, for all* $(x,d) \in \mathcal{E}\times\mathcal{D}$*.*

Assume in addition that the maximum of (5.29) *is attained for an action* $u \in \mathcal{C}(x)$ *and* let us denote by $\pi_T(x,d)$ this action*, then the feedback policy* $\pi = (\pi_{T-k})_{0\le k\le T-1}$ *is an optimal strategy of the problem.*

Sketch of proof. The proof is similar to the one of Theorem 3.20, the differences beeing:

- In Theorem 3.20, there was an additive part which is 0 here.
- In Theorem 3.20, the multiplicative reward R (which was denoted α_k) depends only on the current state Y_k and not on the following state Y_{k+1}, so it was outside the expectation. Thus, one need to adapt the proof.
- Here the process and reward are stationary, so we replaced the backward equation by a forward one, by considering the value as a function of the time remaining until the end.

□

Remark 5.59. When the D_k are independent with law p, then v^T does not depend on d, and we get:

$$v^T(x) = \sup_{u\in\mathcal{C}(x)} \left(\sum_{z\in\mathcal{D}} p_z R(x,u,z) v^{T-1}(f(x,u,z)) \right) \quad \forall x \in \mathcal{E} \ .$$

One can rewrite (5.29) as

$$v^T(x,d) = \sup_{u\in\mathcal{C}(x)} \left(\widetilde{R}(x,d,u) \sum_{(x',z)\in\mathcal{E}\times\mathcal{D}} M^u_{(x,d),(x',z)} v^{T-1}(x',z) \right) \quad \forall (x,d) \in \mathcal{E}\times\mathcal{D} \ ,$$

with

$$\widetilde{R}(x,d,u) = \mathbb{E}\left[R(x,u,D_{k+1}) \mid D_k = d\right] = \sum_{z\in\mathcal{D}} \left(M_{dz} R(x,u,z)\right) \ ,$$

and

$$M^u_{(x,d),(x',z)} := M_{dz} R(x,u,z) \delta_{x'=f(x,u,z)} / \widetilde{R}(x,u,d) \ .$$

Hence, the above finite horizon problem with multiplicative payoff is equivalent to a finite horizon problem with (usual) multiplicative payoff for a MDP on the same state space, and control space but with transition probabilities $M^u_{(x,d),(x',z)}$ and multiplicative reward $\widetilde{R}$.

We have more.

Theorem 5.60. *The above finite horizon problem with multiplicative payoff is equivalent to the finite horizon problem defined for a MDP on the same state space, the control space* $\widetilde{\mathcal{C}}(x) = \mathcal{C}(x) \times \Delta_{\mathcal{E}\times\mathcal{D}}$*, and with transition probabilities* $M^{(u,\theta)}_{yy'} = \theta_{y'}$ *and the following additive payoff criteria*

$$\widetilde{J}^{(T,\sigma)}(y) := J^T(Y; U, \Theta) := \mathbb{E}\left[\left(\sum_{m=0}^{T-1} \tilde{r}(X_m, D_m, U_m, \theta_m)\right) + \widetilde{\varphi}(X_T) \mid Y_0 = y\right] \quad , \tag{5.30}$$

where $\sigma = (\sigma_k)_{k\geq 0}$ *is any strategy in* Σ *or* Σ^{R}*,* $(Y, U, \Theta) := (Y_k, U_k, \theta_k)_{k\geq 0}$ *is the process induced by* σ*, the final reward is* $\widetilde{\varphi} = \log(\varphi)$*, and*

$$\tilde{r}(x, d, u, \theta) = \log \widetilde{R}(x, d, u) - \mathcal{KL}(\theta, M^{(u)}_{(x,d),\cdot}) \tag{5.31}$$

¿ where $\mathcal{KL}$ *is the* Kullback-Leibler distance *(or* entropy*):*

$$\mathcal{KL}(\theta, \theta') = \sum_{(x,d)\in\mathcal{E}\times\mathcal{D}} \theta_{x,d} \log\left(\frac{\theta_{x,d}}{\theta'_{x,d}}\right) \quad .$$

Proof. Take the logarithm of the dynamic programming equation rewritten with $\widetilde{R}$, and denote $\tilde{v}^T(x, d) = \log v^T(x, d)$. Then, $\tilde{v}^T$ satisfies the forward equation (for $(x, d) \in \mathcal{E} \times \mathcal{D}$)

$$\tilde{v}^T(x, d) = \sup_{u\in\mathcal{C}(x)} \left(\log(\widetilde{R}(x, d, u)) + \log\left(\sum_{(x',z)\in\mathcal{E}\times\mathcal{D}} M^u_{(x,d),(x',z)} e^{\tilde{v}^{T-1}(x',z)}\right)\right) \quad .$$

We compare this equation with the dynamic programming equation of the value w^T of the additive criteria problem, which is (for $(x, d) \in \mathcal{E} \times \mathcal{D}$)

$$w^T(x, d) = \sup_{u\in\mathcal{C}(x),\theta\in\Delta_{S\times\mathcal{D}}} \left(\tilde{r}(x, d, u, \theta) + \sum_{(x',z)\in\mathcal{E}\times\mathcal{D}} M^{u,\theta}_{(x,d),(x',z)} w^{T-1}(x', z)\right) \quad .$$

Both equations have the same initial condition $\tilde{v}^0 = \log\varphi = \tilde{\varphi} = w^0$.

So one only need to show, that for all $(x, d) \in \mathcal{E} \times \mathcal{D}$ and $u \in \mathcal{C}(x)$, we have

$$\begin{aligned} &\log(\widetilde{R}(x, d, u)) + \log\left(\sum_{(x',z)\in\mathcal{E}\times\mathcal{D}} M^u_{(x,d),(x',z)} e^{\tilde{v}^{T-1}(x',z)}\right) \\ &= \sup_{\theta\in\Delta_{\mathcal{E}\times\mathcal{D}}} \left(\tilde{r}(x, d, u, \theta) + \sum_{(x',z)\in\mathcal{E}\times\mathcal{D}} M^{u,\theta}_{(x,d),(x',z)} w^{T-1}(x', z)\right) \end{aligned}$$

when $\tilde{v}^{T-1} = w^{T-1}$.

This follows from the following lemma. □

Lemma 5.61. *For any finite set* $\mathcal{E}$*, vector* $v \in \mathbb{R}^{\mathcal{E}}$*, and probability* $\nu \in \Delta_{\mathcal{E}}$*, we have*

$$\log\left(\sum_{x'\in\mathcal{E}} \nu_{x'} e^{v(x')}\right) = \sup_{\theta\in\Delta_S} \left(-\mathcal{KL}(\theta, \nu) + \sum_{x'\in\mathcal{E}} \theta_{x'} v(x')\right)$$

Proof. The map $\psi : v \mapsto \log\left(\sum_{x' \in \mathcal{E}} \nu_{x'} e^{v(x')}\right)$ is convex. This follows from Hölder inequality for the "integral" with respect to ν.

So

$$\psi(v) = \sup_{\theta \in \mathbb{R}^{\mathcal{E}}} \theta \cdot v - \psi^*(\theta)$$

where

$$\psi^*(\theta) = \sup_{v \in \mathbb{R}^{\mathcal{E}}} \theta \cdot v - \psi(v)$$

is the Legendre-Fenchel transform of ψ. Computing ψ^* (by differentiating), we obtain that $\psi^*(\theta) = \mathcal{KL}(\theta, \nu)$ when $\theta \in \Delta_{\mathcal{E}}$ and $+\infty$ otherwise. □

Corollary 5.62. *Consider the dynamic programming equation of the Kullback-Leibler additive criteria:*

$$w^T(x,d) = \sup_{u \in \mathcal{C}(x), \theta \in \Delta_{S \times \mathcal{D}}} \left(\tilde{r}(x,d,u,\theta) + \sum_{(x',z) \in \mathcal{E} \times \mathcal{D}} M^{u,\theta}_{(x,d),(x',z)} w^{T-1}(x',z) \right) \ ,$$

with initial condition $w^0 = \widetilde{\varphi}$. Then, if, for all $(x,d) \in \mathcal{E} \times \mathcal{D}$ and $k \geq 0$, there exists

$$\pi^T(x,d) \in \operatorname*{Argmax}_{u \in \mathcal{C}(x), \theta \in \Delta_{S \times \mathcal{D}}} \left(\tilde{r}(x,d,u,\theta) + \sum_{(x',z) \in \mathcal{E} \times \mathcal{D}} M^{u,\theta}_{(x,d),(x',z)} w^{T-1}(x',z) \right) \ ,$$

and if we denote by $\pi_1^T(x,d)$ the u-coordinate of $\pi^T(x,d)$. Then, π_1^T is an optimal policy for the initial MDP.

5.4.3 Risk sensitive control in infinite horizon

Consider now, for the same MDP and for all strategies $\sigma = (\sigma_k)_{k \geq 0}$ in Σ or Σ^{R}, the *long run time average multiplicative payoffs*:

$$J^{(+,\sigma)}(y) := J^+(Y;U) := \limsup_{T \to \infty} \left(\frac{1}{T} \log \mathbb{E}\left[\prod_{m=0}^{T-1} R(X_m, U_m, D_{m+1}) \mid Y_0 = y \right] \right) \ , \tag{5.32}$$

$$J^{(-,\sigma)}(y) := J^-(Y;U) := \liminf_{T \to \infty} \left(\frac{1}{T} \log \mathbb{E}\left[\prod_{m=0}^{T-1} R(X_m, U_m, D_{m+1}) \mid Y_0 = y \right] \right) \ , \tag{5.33}$$

where $(Y,U) := (Y_k, U_k)_{k \geq 0}$ is the process induced by σ.

Consider the Bellman operator associated to the Kullback-Leibler rewards (given in (5.31)).

This is the operator $\mathcal{B}^{\mathcal{KL}}$ on $\mathbb{R}^{\mathcal{E}\times\mathcal{D}}$ given, for all $(x,d)\in\mathcal{E}\times\mathcal{D}$, by

$$\begin{aligned}\mathcal{B}^{\mathcal{KL}}(w)](x,d) &:= \sup_{u\in\mathcal{C}(x),\theta\in\Delta_{S\times\mathcal{D}}}\left(\tilde{r}(x,d,u,\theta)+\sum_{(x',z)\in\mathcal{E}\times\mathcal{D}}M^{u,\theta}_{(x,d),(x',z)}w^{T-1}(x',z)\right)\\ &= \sup_{u\in\mathcal{C}(x),\theta\in\Delta_{S\times\mathcal{D}}}\left(\log\widetilde{R}(x,d,u)-\mathcal{KL}(\theta,M^{(u)}_{(x,d),\cdot})+\sum_{(x',z)\in\mathcal{E}\times\mathcal{D}}\theta_{(x',z)}w(x',z)\right)\\ &= \sup_{u\in\mathcal{C}(x)}\log\left(\widetilde{R}(x,d,u)\sum_{(x',z)\in\mathcal{E}\times\mathcal{D}}M^{(u)}_{(x,d),(x',d)}e^{w(x',z)}\right)\\ &= \sup_{u\in\mathcal{C}(x)}\left(\sum_{z\in\mathcal{D}}M_{dz}R(x,u,z)e^{w(f(x,u,z),z)}\right)\ .\end{aligned}$$

Theorem 5.63 (The ergodic risk-sensitive dynamic programming equation)**.** *Assume that there exists $\rho\in\mathbb{R}$ and $w\in\mathbb{R}^{\mathcal{E}}$ satisfying the ergodic risk-sensitive dynamic programming equation:*

$$\rho\mathbf{1}+w=\mathcal{B}^{\mathcal{KL}}(w)\ , \tag{5.34}$$

with $\mathcal{B}^{\mathcal{KL}}$ as above. Then, the value function of the long run time average multiplicative Markov decision problem :

$$\zeta^{\pm}(x):=\max_{\sigma}J^{(\pm,\sigma)}(x)\ ,$$

where the maximum is taken over either all relaxed strategies (starting at time 0), or over the restricted sets of pure strategies, Markov strategies, feedback policies, or stationary feedback policies, satisfies

$$\zeta^{\pm}(x)=\rho\quad\forall x\in\mathcal{E}\ .$$

Moreover, if, for all $(x,d)\in\mathcal{E}\times\mathcal{D}$, there exists

$$\pi(x,d)\in\operatorname*{Argmax}_{u\in\mathcal{C}(x),\theta\in\Delta_{S\times\mathcal{D}}}\left(\tilde{r}(x,d,u,\theta)+\sum_{(x',z)\in\mathcal{E}\times\mathcal{D}}\theta_{(x',z)}w(x',z)\right)\ ,$$

and if we denote by $\pi_1(x,d)$ the u-coordinate of $\pi(x,d)$. Then, π_1 is an optimal policy for the long run time average multiplicative MDP.

Note that (5.34) is equivalent to the following equation for $v=e^w$:

$$e^{\rho}v=\sup_{u\in\mathcal{C}(x)}\left(\sum_{z\in\mathcal{D}}M_{dz}R(x,u,z)v(f(x,u,z),z)\right)\ ,$$

and that π_1 can be find directly by using:

$$\pi_1(x,d)\in\operatorname*{Argmax}_{u\in\mathcal{C}(x)}\left(\sum_{z\in\mathcal{D}}M_{dz}R(x,u,z)v(f(x,u,z),z)\right)\ .$$

Consider now the same MDP as above and for all strategies $\sigma = (\sigma_k)_{k\geq 0}$ in Σ or Σ^{R}, the *long run risk sentive payoffs* for $0 < \gamma \leq 1$:

$$J^{(+,\gamma,\sigma)}(y) := J^{+,\gamma}(Y;U) := \limsup_{T\to\infty} \left(\frac{1}{\gamma T} \log \mathbb{E}\left[\prod_{m=0}^{T-1} R^{\gamma}(X_m, U_m, D_{m+1}) \mid Y_0 = y \right] \right) \ , \tag{5.35}$$

$$J^{(-,\gamma,\sigma)}(y) := J^{-,\gamma}(Y;U) := \liminf_{T\to\infty} \left(\frac{1}{\gamma T} \log \mathbb{E}\left[\prod_{m=0}^{T-1} R^{\gamma}(X_m, U_m, D_{m+1}) \mid Y_0 = y \right] \right) \ , \tag{5.36}$$

where $(Y,U) := (Y_k, U_k)_{k\geq 0}$ is the process induced by σ.

The associated Bellman operator is given, for all $(x,d) \in \mathcal{E} \times \mathcal{D}$, by

$$\begin{aligned} \mathcal{B}^{\gamma}(w)](x,d) :=& \sup_{u\in\mathcal{C}(x), \theta\in\Delta_{S\times\mathcal{D}}} \left(\log \widetilde{R}_{\gamma}(x,d,u) - \frac{1}{\gamma}\mathcal{KL}(\theta, M^{(\gamma,u)}_{(x,d),\cdot}) + \sum_{(x',z)\in\mathcal{E}\times\mathcal{D}} \theta_{(x',z)} w(x',z) \right) \\ =& \frac{1}{\gamma} \log \sup_{u\in\mathcal{C}(x)} \left(\sum_{z\in\mathcal{D}} M_{dz} R^{\gamma}(x,u,z) e^{\gamma w(f(x,u,z),z)} \right) \ . \end{aligned}$$

where

$$\widetilde{R}_{\gamma}(x,d,u) = \left(\mathbb{E}\left[R^{\gamma}(x,u,D_{k+1}) \mid D_k = d\right]\right)^{1/\gamma} = \left(\sum_{z\in\mathcal{D}} M_{dz} R(x,u,z) \right)^{\frac{1}{\gamma}} \ ,$$

and

$$M^{\gamma,u}_{(x,d),(x',z)} := M_{dz} R^{\gamma}(x,u,z) \delta_{x'=f(x,u,z)} / (\widetilde{R}_{\gamma}(x,u,d))^{\gamma} \ .$$

Corollary 5.64. *Assume that there exists $\rho^{\gamma} \in \mathbb{R}$ and $w^{\gamma} \in \mathbb{R}^{\mathcal{E}}$ satisfying the ergodic risk-sensitive dynamic programming equation:*

$$\rho^{\gamma}\mathbf{1} + w^{\gamma} = \mathcal{B}^{\gamma}(w^{\gamma}) \ , \tag{5.37}$$

with $\mathcal{B}^{\gamma}$ as above. Then, the value function of the long run risk sentive Markov decision problem :

$$\zeta^{\pm,\gamma}(x) := \max_{\sigma} J^{(\pm,\gamma,\sigma)}(x) \ ,$$

where the maximum is taken over either all relaxed strategies (starting at time 0), or over the restricted sets of pure strategies, Markov strategies, feedback policies, or stationary feedback policies, satisfies

$$\zeta^{\pm,\gamma}(x) = \rho^{\gamma} \quad \forall x \in \mathcal{E} \ .$$

Moreover, if, for all $(x,d) \in \mathcal{E} \times \mathcal{D}$, there exists

$$\pi_1(x,d) \in \operatorname*{Argmax}_{u\in\mathcal{C}(x)} \left(\sum_{z\in\mathcal{D}} M_{dz} R^{\gamma}(x,u,z) e^{\gamma w^{\gamma}(f(x,u,z),z)} \right) \ ,$$

then, π_1 is an optimal policy for the long run risk sentive MDP.

5.5 Problem: Machine replacement

Consider a machine which has several levels of performance depending on its use and that can eventually be repaired/replaced. Let $\mathcal{E} = \{1, \dots, n\}$ be the set of its possible states, where i is a better state than $i+1$, meaning in particular that 1 corresponds to a machine in perfect state. For any given state $i \in S$, we denote by g_i the reward obtained from the use of the machine during an interval of time of one unit (second, minute,...), when it is in state i at the begining of this interval of time and it is not being repaired. We assume that $g_1 > \cdots > g_n$. We denote by R the cost to repair the machine and assume that after reparation during one unit interval of time, the machine is in state 1 at the begining of the next interval. We also assume that, when the machine is in state $i < n$ and is not repaired, then, after one unit of time, it stays in state i with probability $1 - p_i$ and becomes in state $i+1$ with probability p_i. Moreover, if $i = n$ then the machine stays in state n as long as it is not repared. Here, $p_1, \dots, p_n$ are elements of the real interval $(0, 1)$. We would like to optimize the productivity of the machine in the long run.

To solve this problem, we consider a MDP with mean-payoff criterion with for each time $t \in \mathbb{N}$,

X_t : as the state of the machine at the begining of the time interval $[t, t+1)$, where $X_t \in \mathcal{E}$ the state space;

U_t : as the action of the MDP at stage t, equal to 1 if we decide to repare the machine during the time interval $[t, t+1)$, or 0 otherwise, so that the set of actions is $\mathcal{C} = \{0, 1\}$, and we can take also $\mathcal{C}(i) = \mathcal{C}$ for all i (the set of actions is the same for all states of the machine).

The dynamics of the MDP is:

$$P(X_{t+1} = j \mid X_t = i, U_t = u) = M_{i,j}^{(u)}$$

with

$$\begin{aligned}
&M_{i,1}^{(1)} = 1 \quad \forall i \in \mathcal{E} \\
&M_{i,i}^{(0)} = 1 - p_i,\ M_{i,i+1}^{(0)} = p_i, \quad \forall i \in \{1, \dots, n-1\} \\
&M_{n,n}^{(0)} = 1 \\
&M_{i,j}^{(u)} = 0 \quad \text{for all other cases.}
\end{aligned}$$

The reward (gain) at any time t, given the state $x \in \mathcal{E}$ and action $u \in \mathcal{C}$ is

$$g(x; u) = g_x(1 - u) - Ru \ .$$

The payoff of the MDP is the limit when the horizon tends to infinity of the expected productivity in one unit of time, that we try to maximize. Then, the value of the problem is

$$\zeta(i) = \sup \ \limsup_{T \to \infty} \frac{1}{T} \mathbb{E}\left[\sum_{t=0}^{T-1} g(X_t; U_t) \mid X_0 = i \right] \ ,$$

where the supremum is taken among all the feedback strategies defining the process $(U_t)_{t \geq 0}$: $U_t = \pi_t(X_t)$.

Q 5.1. Write the ergodic dynamic programming equation associated to this problem.

Q 5.2. Compute the graph associated to this problem and show that the previous equation has a solution.

Q 5.3. Can we use policy iterations to solve this equation?

Q 5.4. Let v be a bias solution of the ergodic equation, and let π be a stationary feedback control associated to v. Show that $\pi_t = \pi$, $t = 0, \dots, T-1$ is an optimal strategy for the finite horizon problem with reward g and final reward v. What is the interpretation of v and π?

Q 5.5. Consider the finite horizon problem with reward g, final reward $\phi = 0$ and horizon $T = 10$ and assume that $9(g_1 - g_n) < R + g_n$. Show that the optimal strategy is to never repair the machine. Explain the difference with the strategy obtained in the previous case.

Q 5.6. Show that there exists a bias v solution of the ergodic equation, such that v is nonincreasing (that is satisfying $v(x) \geq v(x+1)$ for $x \in \{1, \dots, n-1\}$).

Q 5.7. For $u \in \mathcal{C}$, denote by $F^{(u)}$ the Kolmogorov operator associated to the constant policy $\pi(i) = u$, for all $i \in \mathcal{E}$:

$$F^{(0)}(v)_x = \begin{cases} g_x + (1-p_x)v(x) + p_x v(x+1) & \text{if } x < n \\ g_n + v(n) & \text{if } x = n \end{cases}$$
$$F^{(1)}(v)_x = -R + v(1) \ .$$

Given any nonincreasing bias vector v, show that the optimal policy is to repare the machine when $i > i^*$ only, where

$$i^* = \max\{i \in \mathcal{E} \mid F^{(0)}(v)_i \geq F^{(1)}(v)_i\} \ .$$

Q 5.8. Let ρ, v be a solution of the ergodic equation such that $v(1) = R + \rho$. Show that $v(i) = 0$ for $i > i^*$ and

$$v(i) = R + \rho + \sum_{j=1}^{i-1} \frac{\rho - g_j}{p_j} \quad \text{for } i \leq \min(n, i^*+1) \ .$$

Deduce that the bias is unique up to an additive constant.

Q 5.9. Using the nonincreasing property of v, show that $\rho \leq g_{i^*}$ and that $g_{i^*+1} < \rho$, when $i^* < n$.

Q 5.10. Let

$$w(i) = -g_i + \sum_{j=1}^{i-1} \frac{g_j - g_i}{p_j} \ .$$

Show that w is nondecreasing.

Q 5.11. Show that $w(i^*) \leq R$ and that $R < w(i^*+1)$, when $i^* < n$. Conclude.

5.6 Problem: Portfolio selection with transaction cost

An investor has the choice to invest in a bank account with a fixed proportional return equal to r for each unit of time, or a risky asset (stocks) with a random proportional return equal to α_{k+1} between time k and time $k+1$. The α_k take their values in a finite subset $\mathcal{A}$ of $(0, +\infty)$, and are independent and identically distributed. Transactions on the risky asset induce a proportional cost (commission), which is the same when buying or selling and is denoted by $c \in (0, 1)$.

Q 6.1. Denote by w_k^{b} the amount of money in the bank account and by w_k^{r} the value in money corresponding to what is already invested in the risky asset at the begining of the period of time $[k, k+1]$. Denote also by $w_k = w_k^{\mathrm{b}} + w_k^{\mathrm{r}}$ the total wealth, and by $x_k = \frac{w_k^{\mathrm{r}}}{w_k}$ the proportion of wealth already invested in the risky asset. At the begining of the period $[k, k+1]$, the investor chooses the (new) amount of money $w_{k^+}^{\mathrm{r}}$ to be invested in the risky asset. Let us denote by $u_k \geq 0$ the ratio between this amount and w_k, that is $u_k = \frac{w_{k^+}^{\mathrm{r}}}{w_k}$. Show that the resulting amount of money in the bank account, denoted $w_{k^+}^{\mathrm{b}}$, satisfies:

$$\frac{w_{k^+}^{\mathrm{b}}}{w_k} = 1 - u_k - c|u_k - x_k| \ .$$

Give the conditions on u_k so that $w_{k^+}^{\mathrm{r}}$ and $w_{k^+}^{\mathrm{b}}$ remain nonnegative and write these conditions in the form $u_k \in \mathcal{C}(x_k)$, for some subsets $\mathcal{C}(x)$ of $\mathbb{R}$ to be defined.

Q 6.2. Show that, when $u_k \in \mathcal{C}(x_k)$,

$$\frac{w_{k+1}}{w_k} = R(x_k, u_k, \alpha_{k+1}) \quad \text{and} \quad x_{k+1} = f(x_k, u_k, \alpha_{k+1}),$$

for some maps $R, f : [0,1] \times [0,1] \times \mathbb{R}_+$. Deduce that $(x_k)_{k\geq 0}$ can be seen as a MDP with values in the state space $\mathcal{E} := [0,1]$, controlled by the sequence of actions $(u_k)_{k\geq 0}$ with values in the action space $\mathcal{C} := [0,1]$, and such that $\mathcal{C}(x) \subset \mathcal{C}$ is the possible action space when the state equals x.

Q 6.3. The aim of the investor is to maximize the expected return of the portfolio in the long run. One possible measure of this return is to take the limit when T goes to infinity of $\frac{1}{T}\log\left(\frac{w_T}{w_0}\right)$. Write this problem as an ergodic control problem.

Q 6.4. Write formally the ergodic dynamic programming equation associated to this problem, although the state space is infinite (note that when the sets $\mathcal{C}(x)$ are replaced by finite subsets of $\mathcal{C}(x)$, and x_0 takes its values in a finite subset $\mathcal{E}_0$ of $\mathcal{E}$, the random (finite) sequence $(x_k)_{0\leq k\leq T}$ remains in a finite subset $\mathcal{E}_T$, so that the equation can be shown using the same techniques as in the finite state space case).

Q 6.5. Solve the equation, in the case with no transaction costs, $c = 0$.

Q 6.6. Assume now that the aim of the investor is to maximize the expectation of $\left(\frac{w_T}{w_0}\right)^\gamma$ during a fixed period of time T, where $0 < \gamma \leq 1$ is a parameter. Write formally the dynamic programming equation satisfied by the value v_γ^T of the resulting MDP with finite horizon.

Q 6.7. Write the equation satisfied by $w^T = \frac{1}{\gamma}\log v_\gamma^T$. Show that w^T is the value function of a new MDP with finite horizon and enlarged action space. Which equation computes $\lim_{T\to\infty} w^T/T$?

Q 6.8. Solve the equation, in the case with no transaction costs, $c = 0$.

Additional references for this chapter

[EC1] Serguey Brin and Larry Page. The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30(1-7):107–117, 1998. Proc. 17th International World Wide Web Conference.

[EC2] E. V. Denardo and B. L. Fox. Multichain Markov renewal programs. *SIAM J. Appl. Math.*, 16:468–487, 1968.

[EC3] Tosio Kato. *Perturbation theory for linear operators*. Classics in Mathematics. Springer-Verlag, Berlin, 1995. Reprint of the 1980 edition.

[EC4] E. Seneta. *Nonnegative matrices and Markov chains*. Springer Series in Statistics. Springer-Verlag, New York, second edition, 1981.

Chapter 6

Markov decision problems with partial observation

6.1 Motivation

Until now, we assumed that at each time k, we (the person who decide) know the history of states and actions until this time: $H_k = (X_0, U_0 \ldots, X_{k-1}, U_{k-1}, X_k)$. Then, at each time a decision is taken using the knowledge at that time. This decision consists in choosing an action U_k which is either a deterministic function of the history, $U_k = \sigma_k(H_k) \in \mathcal{C}_k(X_k)$, or a random function of the history, in which case U_k is a random variable with values in $\mathcal{C}_k(X_k)$, and $\sigma_k(H_k)$ is its law. A strategy is then a rule which tells the decision to take at each time, that is a sequence $(\sigma_k)_{k\geq 0}$. The aim is to optimize the expectation of some criteria, or its conditional expectation given the current knowledge, and this optimization is done over all possible strategies.

Now we shall assume that we only know (that is observe or measure) at time k some of the parameters Y_k of the state X_k: Y_k may be (seen as) a projection of the state $X_k \in \mathcal{E}$ and possibly the control $U_{k-1} \in \mathcal{C}_{k-1}$ on a subset $\mathcal{Y}$ of $\mathcal{E}$ or $\mathcal{E}\times\mathcal{C}$, moreover, this projection may be perturbed by noise. We speak of *partially observable Markov decision processes* (POMDP), or *incomplete information 1-players games*. The *information* I_k available at time k, is now the history of observations and actions, that is $I_k = (Y_0, U_0 \ldots, Y_{k-1}, U_{k-1}, Y_k)$. A strategy adapted to the observations now only depend on I_k at time k. For instance a decision associated to a pure strategy satisfies: $U_k = \sigma_k(I_k) \in \mathcal{C}_k$. Since we do not know X_k, we cannot restrict the set of actions by the condition $U_k \in \mathcal{C}_k(X_k)$. One may however consider restrictions of the form $U_k \in \mathcal{C}_k(Y_k)$, or extend the instantaneous rewards by $-\infty$, when $U_k \notin \mathcal{C}_k(Y_k)$.

For instance, for the MDP:

$$X_{n+1} = f_n(X_n, U_n, W_{n+1}), \quad n \geq 0$$

where $X_n \in \mathcal{E}$, $U_n \in \mathcal{C}_n(X_n)$ and $W_n \in \mathcal{W}$, $n \geq 1$, are independent random variables, one may only observe:

$$Y_{n+1} = o_n(X_{n+1}, X_n, U_n, W'_{n+1}) \in \mathcal{Y} \ ,$$

where $W'_n \in \mathcal{W}'$, $n \geq 1$, are independent random variables, independent of the W_k, $k \geq 1$.

The sequence $(W'_n)_{n\geq 1}$ is *the noise on observations*.

Then, one may first ask if we can recover after some time the sequence of states $X_0, \ldots, X_n$ using an appropriate sequence of controls $(U_k)_{k\geq 0}$. At least, we would like to compute and optimize,

given the information at time k, the expectation of

$$\left(\sum_{n=k}^{T} r(X_n, U_n)\right) + \varphi(X_T) \ .$$

Example 6.1. Consider $\mathcal{E} = \mathbb{Z} \cap [-a, a]$, $\mathcal{C} = \{-1, 0, 1\}$, $\mathcal{Y} = \{-1, 0, 1\}$ and the dynamics

$$\begin{aligned} X_{n+1} &= \max(\min(X_n + U_n, a), -a) \\ Y_n &= \mathrm{sgn}(X_n) \quad \text{with } \mathrm{sgn}(0) = 0 \ . \end{aligned}$$

Applying the actions:

$$U_n = -Y_n$$

we obtain

$$X_0 = \tau := \inf\{n \geq 0, Y_n = 0\}$$

and $X_0, \ldots, X_n$ *are recovered*, but this does not allow one to optimize any criteria of the form:

$$\left(\sum_{n=0}^{T} r(X_n, U_n)\right) + \varphi(X_T) \ .$$

Assume now that the dynamics of Y_n is perturbed as in

$$\begin{aligned} X_{n+1} &= \max(\min(X_n + U_n, a), -a) \\ Y_n &= \mathrm{sgn}(X_n + W'_n) \quad \text{with } \mathrm{sgn}(0) = 0 \ , \end{aligned}$$

with $W'_n \in \mathcal{W}' = \{-1, 0, 1\}$ a sequence of independent random variables. Then,

$$X_0 = \tau + 0, \ 1, \text{ or } -1, \text{ where } \tau := \inf\{n \geq 0, Y_n = 0 \text{ or } Y_n Y_{n+1} = -1\}$$

and X_0 is only recovered up to the addition of $0, 1$ or -1.

Example 6.2. Consider $\mathcal{E} = \mathcal{E}_1 \times \mathcal{E}_2$, with $\mathcal{E}_i \subset \mathbb{R}$, $\mathcal{C} = \mathbb{R}_+$, $\mathcal{Y} = \mathcal{E}_1$ and X_n is the couple (position,speed) of a car on a line at time n, U_n is the acceleration assumed to be constant on the interval of time $[n, n+1)$, and Y_n is only the position at time n. Then

$$\begin{aligned} X_{n+1} &= f(X_n, U_n) \\ Y_n &= [X_n]_1 \ , \end{aligned}$$

with $f((x_1, x_2), u) = (u/2 + x_2 + x_1, u + x_2)^T$. This is obtained by integration over the time interval $[n, n+1)$ of the time continuous system:

$$\begin{aligned} \dot{x_1} &= x_2 \\ \dot{x_2} &= u \\ y &= x_1 \ . \end{aligned}$$

Example 6.3. Consider a MDP for which one does not know the model, that is the transition probabilities $M_{xy}^{(u)}$, nor the instantaneous rewards $r(x,u)$ are not known. However, one can observe the sequence of states $X_n \in \mathcal{E}$. Assuming that transition probabilities functions $(x,y,u) \mapsto M_{xy}^{(u)}$ and the reward functions r, belong to a finite set $\mathcal{Z}$, one can consider the enlarged MDP with state space $\mathcal{E} \times \mathcal{Z}$ and state (X_n, Z_n) at time n with $Z_n = z$, for all $n \geq 0$, in which z is the parameter defining the transition probabilities functions $\theta(z)$ and rewards $\eta(z)$, and X_n is the state process of the MDP with transition probabilities $M_{xy}^{(u)} = \theta(z)(x,y,u)$. In that case, one need at the same time to optimize the criterion and learn the model. This is what is done in Q-learning or reinforcement learning.

Example 6.4. Assume that $\mathcal{E} = \{1, 2, \ldots, N\}$ is the set of states of a machine in which larger means worst (N corresponds to breakdown). Also, $\mathcal{C} = \{0, 1, 2\}$ is the set of possible actions: $u = 0$ means that one does not test the machine and thus does not repair it, $u = 1$ means that one test it but does not repair it, and $u = 2$ that one test it and repair it. The set of observations is $\mathcal{Y} = \{0, 1, \ldots, N\}$, where $y = 0$ means that one does not test the machine and so does not know its state, and $y \neq 0$ is the state of the machine when we test it. Then one can consider the following dynamics:

$$X_{n+1} = \begin{cases} 1 & \text{if } U_n = 2, \\ X_n + W_n & \text{if } U_n \leq 1 \end{cases}$$

$$Y_n = \begin{cases} \min(\max(X_n + W_n', 0), N) & \text{if } U_n \geq 1, \\ 0 & \text{otherwise,} \end{cases}$$

where W_n are independent random variables, taking the values 0 and 1, and W_n' are independent random variables, independent of the W_k, taking the values $-1, 0, 1$.

6.2 Partially observable Markov decision processes

Definition 6.5. A *Partially Observable Markov Decision Process (POMDP)* consists in the following parameters:

- a finite or discrete *state space* $\mathcal{E}$;
- a finite or discrete *observation space* $\mathcal{Y}$;
- an *action space* $\mathcal{C}$
- for all $k \in \mathbb{N}$, the subset $\mathcal{C}_k \subset \mathcal{C}$ of all possible actions at time k;
- an initial probability $p^{(0)} \in \Delta_{\mathcal{E}\times\mathcal{Y}}$ on $\mathcal{E} \times \mathcal{Y}$;
- for all $k \in \mathbb{N}$, $x \in \mathcal{E}$ and $u \in \mathcal{C}_k$, a probability row vector $M_x^{(k,u)}$ over $\mathcal{E} \times \mathcal{Y}$, the entries of which will be denoted $\left(M_{xx'}^{(k,u,y')}\right)_{(x',y')\in\mathcal{E}\times\mathcal{Y}}$.

The POMDP is *stationary* if $\mathcal{C}_k$ and $M_x^{(k,u)}$ do not depend on time k. In this case, the index or argument k is omitted. It is *uncontrolled* if the sets $\mathcal{C}_k$ are singletons. In this case, the argument u is omitted.

Formally, a POMDP allows one to construct discrete time processes $(X_k)_{k\geq 0}$, $(Y_k)_{k\geq 0}$ and $(U_k)_{k\geq 0}$ taking their values in $\mathcal{E}$, $\mathcal{Y}$ and $\mathcal{C}$ respectively, with transition probabilities $M_{xx'}^{(k,u,y')}$:

$$M_{xx'}^{(k,u,y')} = P(X_{k+1} = x', Y_{k+1} = y' \mid X_k = x, U_k = u) \ , \tag{6.1a}$$

and such that (X_k, Y_k) satisfies the following Markov property:

$$\begin{aligned} &P\left(X_{k+1} = x_{k+1}, Y_{k+1} = y_{k+1} \mid X_k = x_k, Y_k = y_k, U_k = u_k,\right. \\ &\qquad \left. X_{k-1} = x_{k-1}, \dots, X_0 = x_0, Y_0 = y_0, U_0 = u_0\right) \\ &= P(X_{k+1} = x_{k+1}, Y_{k+1} = y_{k+1} \mid X_k = x_k, U_k = u_k) \ , \\ &\qquad \forall x_i \in \mathcal{E},\ y_i \in \mathcal{Y},\ u_i \in \mathcal{C}_i, \text{with } i \geq 0 \ . \end{aligned} \tag{6.1b}$$

We thus can see $(X_n, Y_n)_{n\geq 0}$ as the process of a MDP but with the additional property that $P(X_{k+1} = x', Y_{k+1} = y' \mid X_k = x, Y_k = y, U_k = u)$ *does not depend on* y.

Definition 6.6. Given a POMDP as above, denote $\mathcal{I}_k = \mathcal{A}_0 \times \cdots \times \mathcal{A}_{k-1} \times \mathcal{Y}$, where (this time) $\mathcal{A}_k := \mathcal{Y} \times \mathcal{C}_k$. This is the set of *informations* at time k.

A *pure strategy* for the POMDP is a sequence $\sigma = (\sigma_k)_{k\geq 0}$ such that, for all $k \geq 0$, σ_k, called the strategy at time k, is a map from $\mathcal{I}_k$ to $\mathcal{C}$ satisfying

$$\sigma_k(i_k) \in \mathcal{C}_k, \text{ for all } i_k \in \mathcal{I}_k \ .$$

We denote by Σ the set of all pure strategies. A pure strategy gives rise to the stochastic process $(X_k, Y_k, U_k, I_k)_{k\geq 0}$ with transition probabilities as in (6.1), satisfying in addition

$$U_k = \sigma_k(I_k), \quad \text{and } I_k = (Y_0, U_0, \dots, Y_{k-1}, U_{k-1}, Y_k) \in \mathcal{I}_k \ ,$$

that is there exists a probability space $(\Omega, \mathfrak{A}, P)$ and a stochastic process $(X_k, Y_k, U_k, I_k)_{k\geq 0}$ over this space satisfying all the above properties.

Such a sequence $(X_k, Y_k, U_k)_{k\geq 0}$ is also called an *admissible sequence* of states, observations, and controls.

Definition 6.7. A *random (or relaxed) strategy* is a sequence $\sigma = (\sigma_k)_{k\geq 0}$ such that, for all $k \geq 0$, σ_k is a map from $\mathcal{I}_k$ to the space of probabilities, denoted here $\mathcal{C}^{\mathrm{R}}$, over a given probability space $(\mathcal{C}, \mathfrak{A}_{\mathcal{C}})$ such that the support of $\sigma_k(i_k)$ is included in $\mathcal{C}_k$, for all $k \geq 0$ and $i_k \in \mathcal{I}_k$.

Such a strategy gives rise to a stochastic process $(X_k, Y_k, U_k, I_k)_{k\geq 0}$ satisfying, for all $B \in \mathfrak{A}_{\mathcal{C}}$,

$$P(U_k \in B \mid I_k) = [\sigma_k(I_k)](B), \quad \text{and } I_k = (Y_0, U_0, \dots, Y_{k-1}, U_{k-1}, Y_k) \in \mathcal{I}_k \ .$$

We denote by Σ^{R} the set of all relaxed strategies.

We can also define the notions of *Markovian* strategies and *feedback* strategies.

However, since $Y_0, \dots, Y_k$ may all be useful to get some information on X_k, the maximum of any criteria over all strategies will not coincide in general with the maximum over all Markovian strategies.

Fact 6.8. Given a POMDP as in Definition 6.5, and a pure strategy $\sigma = (\sigma_k)_{k\geq 0} \in \Sigma$, the associated stochastic process $(X_k, Y_k, U_k, I_k)_{k\geq 0}$ as in Definition 6.6 is such that $(I_k)_{k\geq 0}$ is a Markov chain with initial law $p^{(Y,0)}$:

$$p_y^{(Y,0)} = \sum_{x\in\mathcal{E}} p_{xy}^{(0)} \ .$$

Indeed, since the information I_k is contained in I_{k+1}, and the transition probabilities are:

$$P(I_{k+1} = i_{k+1} \mid I_k = i_k) = \mathcal{M}_{i_k,i_{k+1}}^{(k,\sigma_k(i_k))}$$

where, for $u \in \mathcal{C}_k$,

$$\mathcal{M}_{i_k,i_{k+1}}^{(k,u)} = \begin{cases} P(Y_{k+1} = y_{k+1} \mid I_k = i_k, U_k = u) & \text{if } i_{k+1} = (i_k, u, y_{k+1}) \\ 0 & \text{otherwise.} \end{cases}$$

So (I_k, U_k) is the state-control process of a MDP with state space $\mathcal{I}_k$, control space $\mathcal{C}_k$ at time k, and transition probabilities $\mathcal{M}_{i_k,i_{k+1}}^{(k,u)}$. Similarly, when σ is only a random strategy, then $(I_k, U_k)_{k\geq 0}$ is a Markov chain.

As for MDP, one can consider the following model of POMDP, with a given probability space.

Definition 6.9. A *Partially Observable Markov Decision Process (POMDP)* consists in the following parameters:

- a finite or discrete *state space* $\mathcal{E}$;
- a finite or discrete *observation space* $\mathcal{Y}$;
- an *action space* $\mathcal{C}$;
- for all $k \in \mathbb{N}$, the subset $\mathcal{C}_k \subset \mathcal{C}$ of all possible actions at time k;
- an initial probability $p^{(0)} \in \Delta_{\mathcal{E}\times\mathcal{Y}}$ on $\mathcal{E} \times \mathcal{Y}$;
- a probability space $(\Omega, \mathfrak{A}, P)$, a random variable (X_0, Y_0) with values in $\mathcal{E} \times \mathcal{Y}$ and law $p^{(0)}$, and two sequences of independent random variables $(W_n)_{n\geq 0}$, $(W'_n)_{n\geq 0}$ with values in some discrete spaces $\mathcal{W}$ and $\mathcal{W}'$, independent from each other and independent from (X_0, Y_0);
- for all $k \geq 0$, the *dynamics* of the state at time k, which is a map $f_k : \mathcal{E} \times \mathcal{C}_k \times \mathcal{W} \to \mathcal{E}$, and the dynamics of the *observation*, which is a map $o_k : \mathcal{E} \times \mathcal{E} \times \mathcal{C}_k \times \mathcal{W}' \to \mathcal{Y}$.

The POMDP is *stationary* if $\mathcal{C}_k$, f_k and o_k do not depend on time k, and if the W_k and W'_k are identically distributed. In this case, the index or argument k is omitted. It is *uncontrolled* if the sets $\mathcal{C}_k$ are singletons. In this case, the argument u is omitted.

Given a POMDP in the sense of Definition 6.9, and a strategy of one of the above forms, one can construct on a probability space (extending $(\Omega, \mathfrak{A})$), discrete time processes $(X_k)_{k\geq 0}$, $(Y_k)_{k\geq 0}$ and $(U_k)_{k\geq 0}$ taking their values in $\mathcal{E}$, $\mathcal{Y}$ and $\mathcal{C}$ respectively, satisfying, for all $n \geq 0$:

$$\begin{aligned} X_{n+1} &= f_n(X_n, U_n, W_{n+1}), \\ Y_{n+1} &= o_n(X_{n+1}, X_n, U_n, W'_{n+1}). \end{aligned}$$

Moreover, one can keep the probability space $(\Omega, \mathfrak{A})$ when the strategy is pure.

Fact 6.10. Given a POMDP in the sense of Definition 6.9, we can construct the following transition probabilities which define a POMDP in the sense of Definition 6.5, with same behavior as the initial POMDP:

$$M_{xx'}^{(k,u,y')} = P(f_k(x,u,W_{k+1}) = x',\ o_k(x',x,u,W'_{k+1}) = y')\ .$$

Associated to a Partially Observable Markov decision process, we can consider an optimization problem which consists in maximizing (or minimizing) a criteria equal to the expected value of a functional of the random processes $(X_k)_{k\geq 0}$ and $(U_k)_{k\geq 0}$ induced by the above model among all (relaxed) strategies depending on the information. As for fully observable processes, the criteria can be of several types:

- Finite horizon (time) additive or multiplicative or mixed criteria.
- Infinite horizon discounted (additive) criteria.
- Additive criteria with stopping time, which may be fixed or to be optimized.
- Long run time average criteria.

We shall only discuss here the finite horizon additive criteria.

6.3 POMDP with additive criteria and finite horizon

Let be given a POMDP as in Definition 6.5 or Definition 6.9, and consider or denote:

- for all $k \in \mathbb{N}$, the *instantaneous/running reward/payoff* at time k, which is a map $r_k : \mathcal{E} \times \mathcal{C}_k \to \mathbb{R}$;
- a *final reward*, which is a map $\varphi : \mathcal{E} \to \mathbb{R}$;
- for all strategies $\sigma = (\sigma_k)_{k\geq 0}$ in Σ or Σ^{R} associated to the POMDP, the *total additive payoff* with finite horizon $T \geq 1$:

$$J^{(T,\sigma)} := J^T(X;U) := \mathbb{E}\left[\left(\sum_{k=0}^{T-1} r_k(X_k,U_k)\right) + \varphi(X_T)\right]\ , \tag{6.2}$$

  where $(X,U) := (X_k,U_k)_{k\geq 0}$ is the process induced by σ as in Definition 6.6 or Definition 6.7.
- and the *additive payoff starting at time t with information $i_t \in \mathcal{I}_t$*:

$$J_t^{(T,\sigma)}(i_t) := J_{t,i_t}^T(X;U) := \mathbb{E}\left[\left(\sum_{k=t}^{T-1} r_k(X_k,U_k)\right) + \varphi(X_T) \mid I_t = i_t\right]\ . \tag{6.3}$$

Definition 6.11. A Partially Observable Markov decision problem with the above data consists in the following optimization problem:

$$\max_{\sigma} J^{(T,\sigma)}$$

where the optimization holds over either all relaxed strategies $\sigma \in \Sigma^{\mathrm{R}}$, or all pure strategies associated to the POMDP, that is strategies depending on the information only.

The optimum of above criteria is called the *value* of the problem.

An optimal solution σ is called an *optimal strategy*, and the corresponding process U_k an *optimal control process.*

To compute the optimal strategy, we can rewrite the previous criteria as an additive criteria for the process $(I_n, U_n)_{n\geq 0}$.

Denote (for any strategy σ)

$$\begin{aligned}\tilde{r}_k(i,u) &= \mathbb{E}\left[r_k(X_k,U_k) \mid I_k = i, U_k = u\right] \quad \forall i \in \mathcal{I}_k, u \in \mathcal{C}_k \\ \tilde{\varphi}(i) &= \mathbb{E}\left[\varphi(X_T) \mid I_T = i\right] \quad \forall i \in \mathcal{I}_T \ .\end{aligned}$$

Then,

$$\begin{aligned}\mathbb{E}\left[\tilde{r}_k(I_k,U_k)\right] &= \mathbb{E}\left[r_k(X_k,U_k)\right] \\ \mathbb{E}\left[\tilde{\varphi}(I_T)\right] &= \mathbb{E}\left[\varphi(X_T)\right] \ .\end{aligned}$$

Similarly if we consider the filtration $\mathcal{F}_\ell = \sigma^{\mathrm{a}}(I_\ell, U_\ell)$, then for all $\ell \leq k \leq T$, we have

$$\begin{aligned}\mathbb{E}\left[\tilde{r}_k(I_k,U_k) \mid \mathcal{F}_\ell\right] &= \mathbb{E}\left[r_k(X_k,U_k) \mid \mathcal{F}_\ell\right] \\ \mathbb{E}\left[\tilde{\varphi}(I_T) \mid \mathcal{F}_\ell\right] &= \mathbb{E}\left[\varphi(X_T) \mid \mathcal{F}_\ell\right] \ .\end{aligned}$$

Fact 6.12. For all strategies $\sigma = (\sigma_k)_{k\geq 0}$ in Σ or Σ^{R} associated to the POMDP, the functionals $J^{(T,\sigma)}$ and $J_t^{(T,\sigma)}(i)$ correspond to the additive payoff of the MDP $(I_n, U_n)_{n\geq 0}$, with instanteneous reward $\tilde{r}_k$ at time k, and final reward $\tilde{\varphi}$.

Theorem 6.13 (Dynamic programming equation for POMDP with finite horizon as a function of information)**.** *Assume that the maps φ, r_k, $k \geq 0$ are bounded from above. Let v_k be the value function of the information of the POMDP:*

$$v_k(i_k) := \max_\sigma J_k^{(T,\sigma)}(i_k), \quad \forall i_k \in \mathcal{I}_k \ ,$$

where the maximum is taken over all relaxed strategies starting at time k. Then, v satisfies the following backward recurrence, called the Bellman dynamic programming equation*:*

$$v_k(i_k) = \sup_{u\in\mathcal{C}_k} \left(\tilde{r}_k(i_k,u) + \sum_{i_{k+1}\in\mathcal{I}_{k+1}} \mathcal{M}_{i_k,i_{k+1}}^{(k,u)} v_{k+1}(i_{k+1}) \right) \quad \forall i_k \in \mathcal{I}_k \ . \tag{6.4}$$

with final condition

$$v_T = \tilde{\varphi} \ .$$

Moreover, the values v obtained by optimizing over the restricted set of pure strategies coincide with the one over the set of relaxed strategies.

Assume in addition that the maximum of (6.4) *is attained for an action $u \in \mathcal{C}_k$ and* let us denote by $\sigma_k(i_k)$ this action*, then the pure strategy $\sigma = (\sigma_k)_{0\leq k\leq T-1}$ is an optimal strategy of the problem.*

The difficulty with this result is that $\mathcal{I}_k$ evolves with time k, and that the transition probabilities $\mathcal{M}^{(k,u)}_{i_k,i_{k+1}}$ and the rewards $\tilde{r}_k$ and $\tilde{\varphi}$ are not so easy to compute.

For instance, if $i_k = (y_0, u_0, \dots, y_{k-1}, u_{k-1}, y_k)$, and $i_{k+1} = (i_k, u, y_{k+1})$, then

$$\begin{aligned}\mathcal{M}^{(k,u)}_{i_k,i_{k+1}} &= P(Y_{k+1} = y_{k+1} \mid I_k = i_k, U_k = u) \\ &= \frac{P(I_k = i_k, U_k = u, Y_{k+1} = y_{k+1})}{P(I_k = i_k, U_k = u)}\end{aligned}$$

with for $i_k = (y_0, u_0, \dots, y_{k-1}, u_{k-1}, y_k)$,

$$\begin{aligned}&P(I_k = i_k, U_k = u_k, Y_{k+1} = y_{k+1}) \\ &\quad = P(Y_0 = y_0, U_0 = u_0, \dots, U_k = u_k, Y_{k+1} = y_{k+1}) \\ &\quad = \sum_{x_0,\dots,x_{k+1}} P(Y_0 = y_0, X_0 = x_0, U_0 = u_0, \dots, U_k = u_k, Y_{k+1} = y_{k+1}, X_{k+1} = x_{k+1}) \\ &\quad = \sum_{x_0,\dots,x_{k+1}} p^{(0)}_{x_0,y_0} P(U_0 = u_0 \mid I_0 = i_0) M^{(0,u_0,y_1)}_{x_0,x_1} \cdots P(U_k = u_k \mid I_k = i_k) M^{(k,u_k,y_{k+1})}_{x_k,x_{k+1}}\end{aligned}$$

6.4 A sufficient statistics

Consider the *belief process* $B_k \in \Delta_{\mathcal{E}}$ defined by

$$B_k(x) = P(X_k = x \mid I_k) \quad \forall x \in \mathcal{E} \ .$$

By definition of B_k, it is a measurable function of I_k. Indeed $B_k(x) = \mathbb{E}\left[\mathbf{1}_{X_k=x} \mid I_k\right]$ and

$$B_k = b_k(I_k) \quad \text{with } [b_k(i_k)](x) = P(X_k = x \mid I_k = i_k) = \mathbb{E}\left[B_k(x) \mid I_k = i_k\right] \ \forall i_k \in \mathcal{I}_k \ .$$

Therefore, the set of (pure or random) strategies σ_k which depend only on B_k is smaller than the set of all (pure or random) strategies (which depend on I_k).

We shall show that the optimum of the criteria are the same, or equivalently that the belief process is sufficient to compute the optimal strategy. This is done by proving the following properties for B_k.

Theorem 6.14 (Dynamic programming equation for a sufficient statistics of the POMDP)**.** *Let c_k be (measurable) maps from $\mathcal{I}_k$ to some set $\widetilde{\mathcal{E}}$. Given a strategy $\sigma \in \Sigma^{\mathrm{R}}$, we consider the process $C_k = c_k(I_k)$. Assume that for all $k \geq 0$, $i_k \in \mathcal{I}_k$ and $u_k \in \mathcal{C}_k$, $P(C_{k+1} = c' \mid I_k = i_k, U_k = u_k)$ depends only on $c_k(i_k)$ and u_k, and let us denote by $\widetilde{M}^{(k,u_k)}_{c_k(i_k),c'}$ its value. Assume in addition that $\tilde{r}_k(i_k, u_k)$ depends only on $c_k(i_k)$ and u_k, $\tilde{r}_k(i_k, u_k) = \widetilde{R}_k(c_k(i_k), u_k)$, and $\tilde{\varphi}_T(i_T)$ depends only on $c_T(i_T)$, $\tilde{\varphi}(i_T) = \widetilde{\Phi}(c_T(i_T))$.*

Then, the process (C_k, U_k) defines a MDP with state space $\widetilde{\mathcal{E}}$ and action spaces $\mathcal{C}_k$ and

$$J^{(T,\sigma)} = \mathbb{E}\left[\left(\sum_{k=0}^{T-1} \widetilde{R}_k(C_k, U_k)\right) + \widetilde{\Phi}(C_T)\right] \ .$$

Moreover, assume that the maps φ, r_k, $k \geq 0$ are bounded from above. Then, the value of the POMDP, that is the supremum of the previous functional over all (relaxed) strategies σ of the

POMDP, is equal to the supremum over all strategies of the form $\sigma_k = \sigma'_k \circ c_k$, that is depending on C_k only. More precisely, let $w_k \in \mathbb{R}^{\widetilde{\mathcal{E}}}$ satisfies, for all $c \in \widetilde{\mathcal{E}}$,

$$w_k(c) = \sup_{u \in \mathcal{C}_k} \left(\tilde{R}_k(c,u) + \sum_{c' \in \widetilde{\mathcal{E}}} \widetilde{M}^{(k,u)}_{c,c'} w_{k+1}(c') \right) \quad . \tag{6.5}$$

with final condition

$$w_T = \widetilde{\Phi} \ .$$

Then, w_k is the value function of the MDP (C_k, U_k) with the above criteria, that is, for all $c \in \widetilde{\mathcal{E}}$,

$$w_k(c) = \max_{\sigma} \widetilde{J}^{(T,\sigma)}_k(c) \ ,$$

where, for all $0 \leq t \leq T$,

$$\widetilde{J}^{(T,\sigma)}_t(c) := \mathbb{E}\left[\left(\sum_{k=t}^{T-1} \widetilde{R}_k(C_k, U_k) \right) + \widetilde{\Phi}(C_T) \mid C_t = c \right]$$

and the maximum is taken over all (relaxed) strategies for the process (C_k, U_k), starting at time k.

If v_k is as in Theorem 6.13, then

$$v_k(i_k) = w_k(c_k(i_k)) \quad \forall i_k \in \mathcal{I}_k \ .$$

In particular $v = \mathbb{E}[v_0(I_0)] = \mathbb{E}[w_0(C_0)]$ and the values v obtained by optimizing over the set of pure or relaxed strategies for the POMDP and the restricted set of pure strategies depending only on the C_k coincide.

If, in addition, the maximum in (6.5) is attained for an action $u \in \mathcal{C}_k$ and if we denote by $\sigma'_k(c)$ this action, then the pure strategy $\sigma = (\sigma'_k \circ c_k)_{0 \leq k \leq T-1}$, depending on the C_k, is an optimal strategy of the problem.

When C_k is as above, we say that the process $(C_k)_{k \geq 0}$ is *a sufficient statistics* of the information process $(I_k)_{k \geq 0}$.

Proof. Since the functionals to be maximized for the POMDP and the MDP (C_k, U_k) are the same, it is sufficient to prove that the value functions obtained from the dynamic programming equations are the same, that is to show that $v_k(i_k) = w_k(c_k(i_k))$ for all $i_k \in \mathcal{I}_k$.

Let us show this by backward induction. We have $v_T(i_T) = \tilde{\varphi}(i_T) = \widetilde{\Phi}(c_T(i_T)) = w_T(c_T(i_T))$.

Assume now that $v_{k+1}(i_{k+1}) = w_{k+1}(c_{k+1}(i_{k+1}))$ for all $i_{k+1} \in \mathcal{I}_{k+1}$. Then, for all $i_k \in \mathcal{I}_k$, we have

$$\begin{aligned}
v_k(i_k) &= \sup_{u \in \mathcal{C}_k} \left(\tilde{r}_k(i_k, u) + \sum_{i_{k+1} \in \mathcal{I}_{k+1}} \mathcal{M}^{(k,u)}_{i_k, i_{k+1}} v_{k+1}(i_{k+1}) \right) \\
&= \sup_{u \in \mathcal{C}_k} \left(\widetilde{R}_k(c_k(i_k), u) + \mathbb{E}\left[w_{k+1}(c_{k+1}(I_{k+1})) \mid I_k = i_k, U_k = u \right] \right) \\
&= \sup_{u \in \mathcal{C}_k} \left(\widetilde{R}_k(c_k(i_k), u) + \sum_{c' \in \widetilde{\mathcal{E}}} w_{k+1}(c') P(c_{k+1}(I_{k+1}) = c' \mid I_k = i_k, U_k = u) \right) \\
&= \sup_{u \in \mathcal{C}_k} \left(\widetilde{R}_k(c_k(i_k), u) + \sum_{c' \in \widetilde{\mathcal{E}}} \widetilde{M}^{(k,u)}_{c_k(i_k), c'} w_{k+1}(c') \right) \\
&= w_k(c_k(i_k))
\end{aligned}$$

This shows the induction and finishes the proof. □

The following result shows that B_k satisfies at least the properties relative to the criteria in Theorem 6.14.

Lemma 6.15. *Let $\sigma \in \Sigma$ or Σ^{R} be a strategy for the POMDP, and let $(X_n, Y_n, U_n, I_n)_{n \geq 0}$ be the process induced by σ. Let B_n be the belief process*

$$B_n(x) = P(X_n = x \mid I_n) \quad \forall x \in \mathcal{E} \ .$$

We have

$$\begin{aligned}
\tilde{r}_k(I_k, U_k) &= \sum_{x \in \mathcal{E}} B_k(x) r_k(x, U_k) = R_k(B_k, U_k) \quad \textit{with } R_k(b, u) = b \cdot r_k(\cdot, u) \\
\tilde{\varphi}(I_T) &= \sum_{x \in \mathcal{E}} B_T(x) \varphi(x) = \Phi(B_T) \quad \textit{with } \Phi(b) = b \cdot \varphi \ .
\end{aligned}$$

Therefore

$$J^{(T,\sigma)} = \mathbb{E}\left[\left(\sum_{k=0}^{T-1} R_k(B_k, U_k) \right) + \Phi(B_T) \right] \ .$$

Proof. We have for all $i \in \mathcal{I}_k$ and $u \in \mathcal{C}_k$,

$$\begin{aligned}
\tilde{r}_k(i, u) &= \mathbb{E}\left[r_k(X_k, u) \mid I_k = i, U_k = u \right] \\
&= \sum_{x_k \in \mathcal{E}} r_k(x_k, u) P(X_k = x_k \mid I_k = i, U_k = u)
\end{aligned}$$

Then, using that $U_k = \sigma_k(I_k)$ or has a law equal to $\sigma_k(I_k)$ and is independent of X_k, we get

$$\begin{aligned}
\tilde{r}_k(I_k, U_k) &= \sum_{x_k \in \mathcal{E}} r_k(x_k, U_k) P(X_k = x_k \mid I_k) \\
&= \sum_{x_k \in \mathcal{E}} r_k(x_k, U_k) B_k(x_k) = R_k(B_k, U_k) \ .
\end{aligned}$$

Similarly

$$\tilde{\varphi}(I_T) = \sum_{x \in \mathcal{E}} B_T(x)\varphi(x) \ .$$

□

To prove that $(B_k)_{k\geq 0}$ is a sufficient statistics, and thus use the DP equation of w_k, it remains to show that $P(B_{k+1} = b' \mid I_k = i_k, U_k = u_k)$ depends only on $b_k(i_k)$ and u_k, and thus can be written as $\widetilde{M}^{(k,u_k)}_{b_k(i_k),b'}$.

6.5 The dynamics of the belief process

We first show a formula for the transition probabilities of the information process using the belief process.

Lemma 6.16. *Let $\sigma \in \Sigma$ or Σ^{R} be a strategy for the POMDP, and let $(X_n, Y_n, U_n, I_n)_{n\geq 0}$ be the process induced by σ. Then, the belief process $(B_n)_{n\geq 0}$ satisfies $B_k = b_k(I_k)$ with $b_k(i_k) = \mathbb{E}[B_k \mid I_k = i_k] \in \Delta_{\mathcal{E}}$, for all $k \geq 0$, so it is adapted to the filtration $(\mathcal{F}_n = \sigma^{\mathrm{a}}(I_n))_{n\geq 0}$. Moreover, for all $k \geq 0$, $i_k = (y_0, u_0, \ldots, y_{k-1}, u_{k-1}, y_k)$, and $i_{k+1} = (i_k, u_k, y_{k+1})$, we have*

$$\mathcal{M}^{(k,u_k)}_{i_k,i_{k+1}} = \sum_{x_k, x_{k+1} \in \mathcal{E}} \left([b_k(i_k)](x_k) M^{(k,u_k,y_{k+1})}_{x_k,x_{k+1}} \right) = b_k(i_k) M^{(k,u_k,y_{k+1})} \mathbf{1} \ .$$

Proof. Let $k \geq 0$, $i_k = (y_0, u_0, \ldots, y_{k-1}, u_{k-1}, y_k)$, and $i_{k+1} = (i_k, u_k, y_{k+1})$. We have

$$\begin{aligned} \mathcal{M}^{(k,u_k)}_{i_k,i_{k+1}} &= P(Y_{k+1} = y_{k+1} \mid I_k = i_k, U_k = u_k) \\ &= \frac{P(I_k = i_k, U_k = u_k, Y_{k+1} = y_{k+1})}{P(I_k = i_k, U_k = u_k)} \end{aligned}$$

with

$$\begin{aligned} &P(I_k = i_k, U_k = u_k, Y_{k+1} = y_{k+1}) \\ &\quad = \sum_{x_k, x_{k+1}} P(I_k = i_k, U_k = u_k, X_k = x_k, Y_{k+1} = y_{k+1}, X_{k+1} = x_{k+1}) \\ &\quad = \sum_{x_k, x_{k+1}} P(I_k = i_k, X_k = x_k) P(U_k = u_k \mid I_k = i_k) M^{(k,u_k,y_{k+1})}_{x_k,x_{k+1}} \\ &\quad = \sum_{x_k, x_{k+1}} P(I_k = i_k) P(X_k = x_k \mid I_k = i_k) P(U_k = u_k \mid I_k = i_k) M^{(k,u_k,y_{k+1})}_{x_k,x_{k+1}} \\ &\quad = \sum_{x_k, x_{k+1}} P(I_k = i_k, U_k = u_k) P(X_k = x_k \mid I_k = i_k) M^{(k,u_k,y_{k+1})}_{x_k,x_{k+1}} \end{aligned}$$

Hence

$$\mathcal{M}^{(k,u)}_{i_k,i_{k+1}} = \sum_{x_k, x_{k+1} \in \mathcal{E}} P(X_k = x_k \mid I_k = i_k) M^{(k,u_k,y_{k+1})}_{x_k,x_{k+1}} \ ,$$

and since $P(X_k = x_k \mid I_k = i_k) = \mathbb{E}[B_k(x_k) \mid I_k = i_k] = [b_k(i_k)](x_k)$, we get the result. □

Corollary 6.17. *We have, for all $i_k \in \mathcal{I}_k$, $u_k \in \mathcal{C}_k$, and $b' \in \Delta_{\mathcal{E}}$,*

$$P(B_{k+1} = b' \mid I_k = i_k, U_k = u_k) = \sum_{y_{k+1} \in \mathcal{Y}, b_{k+1}(i_k, u_k, y_{k+1}) = b'} b_k(i_k) M^{(k,u_k,y_{k+1})} \mathbf{1} \ .$$

Proof. Use

$$\begin{aligned} P(B_{k+1} = b' \mid I_k = i_k, U_k = u_k) &= P(b_{k+1}(I_{k+1}) = b' \mid I_k = i_k, U_k = u_k) \\ &= \sum_{i_{k+1} \in \mathcal{I}_{k+1},\ b_{k+1}(i_{k+1}) = b'} \mathcal{M}^{(k,u_k)}_{i_k, i_{k+1}} \ . \end{aligned}$$

□

In view of previous formula, it is now sufficient to show that $b_{k+1}(i_k, u_k, y_{k+1})$ only depends on $b_k(i_k)$.

Let us first show the property in the case with no control. In this case, $M^{(k,y')}$ is the matrix with entries $M^{(k,y')}_{xx'}$, $x, x' \in \mathcal{E}$. We shall use the notation

$$\mathcal{N}(b) = \frac{1}{b\mathbf{1}} b \in \Delta_{\mathcal{E}}, \quad \forall b \in \mathbb{R}_+^{\mathcal{E}} \ .$$

This is a normalization: $\mathcal{N}(\lambda b) = \mathcal{N}(b)$, for all $\lambda \in \mathbb{R}_+$.

Proposition 6.18. *If $I_k = (Y_0, \ldots, Y_k)$, $i_k = (y_0, \ldots, y_k)$ and $[b_k(i_k)](x) = P(X_k = x \mid I_k = i_k)$, then*

$$b_{k+1}(i_k, y_{k+1}) = \mathcal{N}(b_k(i_k) M^{(k,y_{k+1})}) \quad \text{and} \quad b_0(i_0) = \mathcal{N}(p^{(0)}_{\cdot, i_0}) \ .$$

Proof. Let us fix i_k and denote $q^{(k)}_x = P(X_k = x, I_k = i_k)$, and similarly for i_{k+1} and $q^{(k+1)}$. Since $[b_k(i_k)](x) = P(X_k = x \mid I_k = i_k)$, we get $b_k(i_k) = \mathcal{N}(q^{(k)})$.

We have

$$\begin{aligned} q^{(k+1)}_x =& P(X_{k+1} = x, Y_{k+1} = y_{k+1}, I_k = i_k) \\ &= \sum_{x' \in \mathcal{E}} P(X_{k+1} = x, Y_{k+1} = y_{k+1}, X_k = x', I_k = i_k) \\ &= \sum_{x' \in \mathcal{E}} P(X_{k+1} = x, Y_{k+1} = y_{k+1} \mid X_k = x') P(X_k = x', I_k = i_k) \\ &= \sum_{x' \in \mathcal{E}} M^{(k,y_{k+1})}_{x'x} q^{(k)}_{x'} = [q^{(k)} M^{(k,y_{k+1})}]_x \end{aligned}$$

Therefore $q^{(k+1)} = q^{(k)} M^{(k,y_{k+1})}$, hence

$$b_{k+1}(i_{k+1}) = \mathcal{N}(q^{(k+1)}) = \mathcal{N}(q^{(k)} M^{(k,y_{k+1})}) \ ,$$

and since $b_k(i_k) = \mathcal{N}(q^{(k)})$ is proportional to $q^{(k)}$, we deduce $b_{k+1}(i_{k+1}) = \mathcal{N}(b_k(i_k) M^{(k,y_{k+1})})$. □

Generalizing the previous result to the controlled case, we obtain:

Corollary 6.19. *Given a strategy $\sigma \in \Sigma$, with $\sigma_k : \mathcal{I}_k \to \mathcal{C}_k$, for all $k \geq 0$, we have for all $k \geq 0$, $u_k \in \mathcal{C}_k$, $y_0, y_{k+1} \in \mathcal{Y}$ and $i_k \in \mathcal{I}_k$*

$$b_{k+1}(i_k, u_k, y_{k+1}) = \mathcal{N}(b_k(i_k) M^{(k,u_k,y_{k+1})}) \quad \text{and} \quad b_0(y_0) = \mathcal{N}(p^{(0)}_{\cdot, y_0}) \ .$$

Therefore the dynamics of the belief process is given by:

$$B_{k+1} = \mathcal{N}(B_k M^{(k,U_k,Y_{k+1})}) \ .$$

Note first that the above dynamics is linear up to a normalization. Moreover, the future observation Y_{k+1} plays the role of a random perturbation of the dynamics of the state process B_k.

Corollary 6.20. *For all $b' \in \Delta_{\mathcal{E}}$, we have that*

$$P(B_{k+1} = b' \mid I_k = i_k, U_k = u_k) = \sum_{y_{k+1} \in \mathcal{Y}, \mathcal{N}(b_k(i_k) M^{(k,u_k,y_{k+1})}) = b'} b_k(i_k) M^{(k,u_k,y_{k+1})} \mathbf{1}$$

depends only on $b_k(i_k)$ and u_k, and thus can be written as $\widetilde{M}^{(k,u_k)}_{b_k(i_k),b'}$ with

$$\widetilde{M}^{(k,u)}_{b,b'} = \sum_{y' \in \mathcal{Y}, \mathcal{N}(bM^{(k,u,y')}) = b'} \left(bM^{(k,u,y')} \mathbf{1} \right)$$

This shows that the belief process B_k is a sufficient statistics for the information process I_k.

The Dynamic programming equation associated to the belief process B_k is

$$w_k(b) = \sup_{u \in \mathcal{C}_k} \left(R_k(b,u) + \sum_{b' \in \Delta_{\mathcal{E}}} \widetilde{M}^{(k,u)}_{b,b'} w_{k+1}(b') \right)$$

with final condition

$$w_T = \Phi \ ,$$

in which

$$\begin{aligned} R_k(b,u) &= b \cdot r_k(\cdot, u) \\ \Phi(b) &= b \cdot \varphi \\ \widetilde{M}^{(k,u)}_{b,b'} &= \sum_{y' \in \mathcal{Y}, \mathcal{N}(bM^{(k,u,y')}) = b'} \left(bM^{(k,u,y')} \mathbf{1} \right) \ . \end{aligned}$$

It can then be rewritten as:

$$w_k(b) = \sup_{u \in \mathcal{C}_k} \left(b \cdot r_k(\cdot, u) + \sum_{y' \in \mathcal{Y}} \left((bM^{(k,u,y')} \mathbf{1}) w_{k+1}(\mathcal{N}(bM^{(k,u,y')})) \right) \right) \ . \tag{6.6}$$

So the sequence $(Y_k)_{k \geq 0}$ is like a sequence of independent random variables, such that the law of Y_{k+1} depends on B_k.

Theorem 6.21 (Dynamic programming equation for the POMDP as a function of belief)**.** *Assume that the maps φ, r_k, $k \geq 0$ are bounded from above. Then, the value of the POMDP is equal to the value of the MDP for the belief process:*

$$B_{k+1} = \mathcal{N}(B_k M^{(k,U_k,Y_{k+1})}) \ ,$$

starting at $B_0 = \mathcal{N}(p^{(0)}_{\cdot,Y_0})$. If w_k is solution of (6.6) *with the final condition $w_T(b) = b \cdot \varphi$, then $v = \mathbb{E}[w_0(B_0)]$.*

Assume in addition that the maximum in (6.6) *is attained for an action $u \in \mathcal{C}_k$ and* let us denote by $\sigma_k(b)$ this action, *then the pure strategy $\sigma = (\sigma_k)_{0 \leq k \leq T-1}$, depending on the B_k, is an optimal strategy of the problem.*

- If $\mathcal{E}$, $\mathcal{Y}$ and $\mathcal{C}_k$ are finite, then the sets $\mathcal{I}_k$ are finite and so the set of possible values of B_k in $\Delta_{\mathcal{E}}$ is also finite. To be in the case of a MDP with finite state space, we still need to consider a variable state space: take $\mathcal{P}_k$ be the set of possible values of B_k.
- One can show that w_k is a convex function of $b \in \Delta_{\mathcal{E}}$.
- One can show that w_k is Lipschitz continuous with a constant less than $\sum_{\ell=k}^{T-1} \|r_\ell\|_\infty + \|\varphi\|_\infty$.

6.6 Infinite horizon problems

One can generalize the above results to discounted infinite horizon criteria. In that case, the belief MDP belongs to an infinite state space. However, the set of possible values of B_k can still be restricted to a countable set, by taking the reachable set from B_0.

6.7 Problem: Machine replacement with partial observation

Consider a machine which has two levels of performance: working (1) or breakdown (0). When it is working, it can return R euros by time unit. When it is broken, it returns nothing. We do not have access to the state of the machine, however we can choose to do some test which gives its state. The test costs C_t euros and repairing the machine (if it is broken) costs C_r euros. We denote

X_k : the state of the machine at time k (at the begining of the time period $[k, k+1)$), $X_k = 0$ if the machine is broken, and $X_k = 1$ if it works;

U_k : the action taken at time k, $U_k = 1$ means that we test the machine and we repair it it is broken. and $U_k = 0$ means that we do not test it;

Y_k : the information we have on the state of the machine at time $k-1$ in case we have tested it: $Y_k = X_{k-1}$ if $U_{k-1} = 1$ and $Y_k = -1$ otherwise.

We assume that if the machine is working at time k, it will still work at time $k+1$ with probability p, and it will be broken at time $k+1$ with probability $1-p$.

6.7.1 The corresponding POMDP

Q 7.1. Construct a MDP satisfying the above properties for X_k with U_k equal to the decision to repair or not: give the transition probabilities.

Q 7.2. Construct a POMDP satisfying the above properties, with Y_k as the observation, and U_k as the action. Show that (X_k, Y_k) is the sequence of states of a MDP, the transition probabilities of which do not depend on past of observations Y_k. Determine the transition probabilities:

$$M_{xx'}^{(u,y')} := P(X_{k+1} = x', Y_{k+1} = y' | X_k = x, Y_k = y, U_k = u) \ .$$

Q 7.3. The aim is to maximize the sum of the rewards during a period of T time units. Write this problem as a POMDP with finite horizon. Precise the parameters of the problem.

Q 7.4. Write the dynamic programming equation satisfied by the value function v_k^T for the criteria with finite horizon starting at time k, as a function of the law $q = (q_0, q_1) \in \Delta_{\mathcal{E}}$ of X_k given the past information.

Q 7.5. Show that $v_k^T = v_0^{T-k}$, and that $v_0^T(q) = z_T(q_1)$, for some map $z_T : [0,1] \to \mathbb{R}$ which satisfies the recurrence:

$$z_{T+1} = \mathcal{B}(z_T)$$

with initial condition $z_0 = 0$, where $\mathcal{B} : \mathbb{R}^{[0,1]} \to \mathbb{R}^{[0,1]}$ satisfies:

$$[\mathcal{B}(z)](q) = \max([\mathcal{B}^{(0)}(z)](q), [\mathcal{B}^{(1)}(z)](q))$$

with

$$[\mathcal{B}^{(0)}(z)](q) = Rq + z(qp)$$
$$[\mathcal{B}^{(1)}(z)](q) = -(C_t + C_r) + (R + C_r)q + (1-q)z(1) + qz(p) \ .$$

Gives the optimal policy $\pi_T(q)$ by using this equation.

6.7.2 Solving the DP equation

Q 7.6. Show that $\mathcal{B}^{(0)}, \mathcal{B}^{(1)}$ and thus $\mathcal{B}$ preserve the set $\mathcal{C}$ of convex function that are piecewise affine. Deduce that $z_k \in \mathcal{C}$ for all $k \geq 0$.

Q 7.7. Show that $z_1 \geq z_0$ and deduce that $z_{k+1} \geq z_k$.

Q 7.8. Let $a = \max(R + C_r, R/(1-p))$. Show that $\mathcal{B}$ preserves the set of functions v that are piecewise differentiable with a derivative in $[0, a]$. Deduce that z_k is a nondecreasing function, for all $k \geq 0$.

Q 7.9. denote $\varphi_{k+1} = \mathcal{B}^{(0)}(z_k) - \mathcal{B}^{(1)}(z_k)$. Show that $\varphi_{k+1}(1) > 0$ and that

$$\varphi_{k+1}(0) \leq \max(\varphi_k(0), 0) \ .$$

Q 7.10. Let $k^* := \min\{k \geq 0 \mid \varphi_k(0) \leq 0\}$. Show that for all $k \geq k^*$, there exist $0 \leq s_k < 1$ such that $\varphi_k(x) > 0 \Leftrightarrow x > s_k$. Deduce that an optimal policy can be obtained such that $\pi_k(x) = 0$ for $x \geq s_k$ and $\pi_k(x) = 1$ for $x < s_k$.

Q 7.11. Show by induction that for all $k < k^*$, z_k and φ_{k+1} are affines, and that φ_{k+1} is positive on all $[0, 1]$. Deduce that the threshold $s_k = 0$ is possible for $k \leq k^*$.

Chapter 7

Constrained Markov decision processes and Linear programming formulation of Dynamic programming

7.1 Motivation

Example 7.1 (Dam optimization)**.** Consider the management of a hydroelectric dam subject to a tourist constraint: one optimizes the revenue of the electrical production with the constraint that the dam has a minimal level with a given tolerance level in probability. Such a problem is often called a *chance constrained control/optimization problem.*

For instance, let $b \in (0,1)$ be the probability level, $\mathcal{T}$ be a subset of $\{0,1,\dots,T-1\}$ representing the *tourist season*, and $\psi(x)$ be the dam level. The dam optimization can be as follows:

$$\begin{aligned}&\max \mathbb{E}\left[\left(\sum_{k=0}^{T-1} r(X_k,U_k)\right)+\varphi(X_T)\right]\\ &\text{under the constraint } P(\psi(X_k)\geq a,\ \forall k\in\mathcal{T})\geq b\\ &\text{and } X_{k+1}=f_k(X_k,U_k,W_k)\ .\end{aligned}$$

with $(W_k)_{k\geq 0}$ a sequence of independent random variables.

Since $P(X\geq a)=\mathbb{E}\left[\mathbf{1}_{X\geq a}\right]$, we can rewrite the constraint as a constraint on a functional of the same type as the one to be optimized (although it contains only a final reward):

$$P(\psi(X_k)\geq a,\ \forall k\in\mathcal{T})=\mathbb{E}\left[Z_T\right]$$

where $Z_0=1$ and for all $k\geq 0$,

$$Z_{k+1}=\begin{cases}Z_k & \text{if } \psi(X_k)\geq a \text{ or } k\notin\mathcal{T}\\ 0 & \text{otherwise.}\end{cases}$$

Then, taking (X_k,Z_k) as the new state at time k, we obtain a *constrained MDP with finite horizon additive crirerion and constraints.*

Example 7.2 (Pagerank Optimization)**.** Consider the pagerank optimization:

$$\max_{P\in\mathcal{P}} \sum_{x\in W} g(x)p_x^M \ ,$$

in which

- $\mathcal{E}$ is the set of *Web pages*;
- $W \subset \mathcal{E}$ is the web site to be optimized;
- $\mathcal{P}$ is a set of possible $\mathcal{E}\times\mathcal{E}$ Markov matrices, like Markov transition matrices of simple random walks on the possible Web graphs;
- $M = \gamma P + (1-\gamma)\mathbf{1}z$, where $0 < \gamma < 1$ is the *damping factor*;
- $g \in \mathbb{R}_+^W$ is a vector of weights.

When $\mathcal{P}$ is local, meaning that

$$\mathcal{P} = \{P \in \mathbb{R}_+^{\mathcal{E}\times\mathcal{E}} \mid P_{x\cdot} \in \mathcal{C}(x)\} \ ,$$

we can reduce this problem to a long run time average payoff problem for a MDP, see Section 5.3.2.

However, *frequency constraints* such as

$$P(X_{k+1} \in J \mid X_k \in I) \leq b \tag{7.1}$$

with $b \in (0,1)$ cannot be put into a *local* constraint of the form $\mathcal{C}(x)$.

We have

$$P(X_{k+1} \in J \mid X_k \in I) = \frac{\sum_{x\in I,\ y\in J} P(X_k = x,\ X_{k+1} = y)}{\sum_{x\in I,\ y\in\mathcal{E}} P(X_k = x,\ X_{k+1} = y)} = \frac{\sum_{x\in I,\ y\in J} p_x^M M_{xy}}{\sum_{x\in I,\ y\in\mathcal{E}} p_x^M M_{xy}} \ .$$

The frequency constraint (7.1) is then equivalent to

$$\sum_{x\in I,\ y\in J} p_x^M M_{xy} \leq b \sum_{x\in I,\ y\in\mathcal{E}} p_x^M M_{xy} \ .$$

and so can be put in the form:

$$\sum_{x\in\mathcal{E},\ y\in\mathcal{E}} h(x,y)p_x^M M_{xy} \leq 0$$

with

$$h(x,y) = \begin{cases} 1-b & \text{if } x \in I,\ y \in J \\ -b & \text{if } x \in I,\ y \in \mathcal{E}\setminus J \\ 0 & \text{otherwise } (x \in \mathcal{E}\setminus I). \end{cases}$$

Since $M = \gamma P + (1-\gamma)\mathbf{1}z$, and the row x of P, $P_{x\cdot}$, is the action $\pi(x)$ choosen in state x, we can consider

$$\tilde{h}(x,u) = \sum_{y\in\mathcal{E}} (h(x,y)(\gamma u_y + (1-\gamma)z_y)) \ .$$

Then, the constraint is equivalent to the following constraint on the policy π:

$$\sum_{x\in\mathcal{E}}\left(\tilde{h}(x,\pi(x))p_x^{(\pi)}\right)\leq 0$$

which is equivalent to

$$G^{(\pm,\pi)}(x):=\lim_{T\to\infty}\left\{\frac{1}{T}\mathbb{E}\left[\sum_{k=0}^{T}h(X_k,U_k)\mid X_0=x\right]\right\}\leq 0.$$

We then obtain a *constrained MDP with long run time average payoff and constraints.*

7.2 Constrained MDP with finite horizon

Let be given a Markov decision process as in Definition 3.1, that is the following parameters:

- a finite or discrete *state space* $\mathcal{E}$;
- an *action space* $\mathcal{C}$
- for all $k\in\mathbb{N}$ and $x\in\mathcal{E}$, the subset $\mathcal{C}_k(x)\subset\mathcal{C}$ of all possible actions at time k, when the state is equal to x;
- for all $k\in\mathbb{N}$, the set $\mathcal{A}_k:=\{(x,u)\mid x\in\mathcal{E},\ u\in\mathcal{C}_k(x)\}$ of all possibles couples (state, action) at time k;
- an initial probability $p^{(0)}\in\Delta_{\mathcal{E}}$ on $\mathcal{E}$, or an initial state $x_0\in\mathcal{E}$, which is equivalent to the case where $p^{(0)}$ is the Dirac measure at x_0;
- for all $k\in\mathbb{N}$, $x\in\mathcal{E}$ and $u\in\mathcal{C}_k(x)$, a probability row vector $M_x^{(k,u)}$ over $\mathcal{E}$, the entries of which will be denoted $\left(M_{xy}^{(k,u)}\right)_{y\in\mathcal{E}}$.

One can alternatively consider a MDP in the sense of Definition 3.8. Consider several instantaneous and final rewards: some will be used in the functional to be optimized, the other ones will be used in the constraint functionals:

- for all $k\in\mathbb{N}$, the *instantaneous rewards* at time k, are maps r_k and g_k^ℓ, $1\leq\ell\leq L$, $\mathcal{A}_k\to\mathbb{R}$;
- the *final rewards* are maps φ, and ψ^ℓ, $1\leq\ell\leq L$, $\mathcal{E}\to\mathbb{R}$;

For all strategies $\sigma=(\sigma_k)_{k\geq 0}$ in Σ or Σ^{R} (or Π or Π^{R}), consider the *total additive payoffs* with finite horizon $T\geq 1$:

$$J^{(T,\sigma)}:=J^T(X;U):=\mathbb{E}\left[\left(\sum_{k=0}^{T-1}r_k(X_k,U_k)\right)+\varphi(X_T)\right] \tag{7.2a}$$

$$G^{(\ell,T,\sigma)}:=J^{\ell,T}(X;U):=\mathbb{E}\left[\left(\sum_{k=0}^{T-1}g_k^\ell(X_k,U_k)\right)+\psi^\ell(X_T)\right]\ , \tag{7.2b}$$

where $(X,U):=(X_k,U_k)_{k\geq 0}$ is the process induced by σ (as in Definition 3.2 or Definition 3.3).

Definition 7.3. A constrained Markov decision problem with complete observation, and finite horizon consists in the following optimization problem:

$$\max_{\sigma} J^{(T,\sigma)} \text{ under the constraints } G^{(\ell,T,\sigma)} \leq h_\ell \quad \forall 1 \leq \ell \leq L \ ,$$

where the optimization holds over either all relaxed strategies $\sigma \in \Sigma^{\mathrm{R}}$, or all pure strategies, or all Markov strategies, or all feedback policies, and where $h_l \in \mathbb{R}$, $1 \leq \ell \leq L$, are some given thresholds.

The optimum of above criteria is called the *value* of the problem. An optimal solution σ is called an *optimal strategy*, and the corresponding process U_k or (X_k, U_k) an *optimal control process.*

Contrarily to the unconstrained case we have

Fact 7.4. For a constrained MDP, the value over the set of random strategies may be different from the value over the set of pure strategies.

Example 7.5 (A counter-example)**.** Consider the simplest case: $T = 1$, $\mathcal{E} = \{1\}$, $\mathcal{C} = \{0, 1\}$, $\ell = 1$ (one constraint). Since $T = 1$, we only need one reward function r_0, strategies coincide with Markovian strategies and consist in one policy at time 0, so we shall ommit the index 0 in r_0 and g_0 and denote the policy at time 0 by π. Since $\mathcal{E} = \{1\}$, we omit the parameter x, and policies are probabilities on $\mathcal{C}$, or equivalently probability row vectors $\pi = (\pi_0, \pi_1)$. Pure strategies correspond to the Dirac probabilities in 0 or 1. Then

$$\Sigma^{\mathrm{R}} = \{(\pi_0, \pi_1) \in \mathbb{R}^2 \mid \pi_0 + \pi_1 = 1,\ \pi_0 \in [0, 1]\}, \quad \Sigma = \{(0, 1), (1, 0)\}.$$

Denote by $h = h_1$ the threshold of the constraint, and by $v(h)$ *the value over the set of pure strategies* and $v^{\mathrm{R}}(h)$ *the value over the set of random strategies.*

We have

$$v^{\mathrm{R}}(h) = \max\{r(0)\pi_0 + r(1)\pi_1 + \varphi \mid \pi \in \Sigma^{\mathrm{R}},\ g(0)\pi_0 + g(1)\pi_1 + \psi \leq h\}$$

and

$$v(h) = \max\{r(0)\pi_0 + r(1)\pi_1 + \varphi \mid \pi \in \Sigma,\ g(0)\pi_0 + g(1)\pi_1 + \psi \leq h\} \ .$$

The first problem is the maximization of an affine function over the convex subset of Σ^{R} satisfying the linear inequality constraint. The optimum is attained on extremal points of this subset.

With no constraints, the set of extremal points is Σ. In general it is not.

Choose

$$r(0) = 1,\ r(1) = 0,\ g(0) = 1,\ g(1) = 0,\ \varphi = \psi = 0 \ .$$

Then

$$v^{\mathrm{R}}(h) = \max\{\pi_0 \mid \pi_0 \in [0, 1],\ \pi_0 \leq h\}$$

and

$$v(h) = \max\{\pi_0 \mid \pi_0 \in \{0, 1\},\ \pi_0 \leq h\} \ .$$

We then obtain

$$v^{\mathrm{R}}(h) = \begin{cases} -\infty & \text{if } h < 0 \qquad \text{no solution} \\ h & \text{if } 0 \leq h < 1 \qquad \pi_0 = h \\ 1 & \text{if } h \geq 1 \qquad \pi_0 = 1 \ . \end{cases}$$

and

$$v(h) = \begin{cases} -\infty & \text{if } h < 0 \qquad \text{no solution} \\ 0 & \text{if } 0 \le h < 1 \qquad \pi_0 = 0 \\ 1 & \text{if } h \ge 1 \qquad \pi_0 = 1 \ . \end{cases}$$

Hence

$$v(h) < v^{\mathrm{R}}(h) \quad \text{for } h \in (0,1) \ .$$

We already saw in this simple example that the optimization over random strategies is a *Linear Program*, that is the optimization of an affine functional over a subset K of vectors satisfying linear inequality constraints. Moreover, the optimization over pure strategies corresponds to the same criteria but on a subset of K. Then, the optimization over random strategies consists in a *relaxation* or *convexification* of the optimization over pure strategies.

Some attempts (see works of Chen and Blankenship [CS2]) have been done to solve a constrained MDP over the set of *pure feedback strategies*, by using a dynamic programming approach, but this is to the prize of increasing the state space with the threshold parameters h_ℓ, and the control space with set of functions.

To solve a constrained MDP, over the set of random feedback or Markovian strategies, we shall rewrite the optimization as a Linear Program on an appropriate set of vectors. Instead of a random policy $\pi \in \Pi^{\mathrm{R}}$, which gives for all k and x, the probability law of U_k given $X_k = x$, we shall consider the *occupation measure* which is the induced probability law of (X_k, U_k).

Let $\pi = (\pi_k)_{k \ge 0} \in \Pi^{\mathrm{R}}$ be a random Markovian strategy, and denote $\pi_k(\cdot \mid x)$ the probability measure $\pi_k(x)$. For all $k \ge 0$, the *associated Markov matrix and reward vectors at time k* will be denoted $M^{(k,\pi)}$, $r^{(k,\pi)}$, $g^{(k,\ell,\pi)}$, for $1 \le \ell \le L$, and are given, for $x \in \mathcal{E}$, by

$$\begin{aligned} r_x^{(k,\pi)} &:= \int_{\mathcal{C}} r_k(x,u)\pi_k(du \mid x) \\ g_x^{(k,\ell,\pi)} &:= \int_{\mathcal{C}} g_k^\ell(x,u)\pi_k(du \mid x) \\ M_{xy}^{(k,\pi)} &:= \int_{\mathcal{C}} M_{xy}^{(k,u)}\pi_k(du \mid x) \ . \end{aligned}$$

If $(X,U) := (X_k, U_k)_{k \ge 0}$ is the process induced by π, we get

$$\begin{aligned} r_x^{(k,\pi)} &= \mathbb{E}\left[r_k(X_k,U_k) \mid X_k = x\right] \\ g_x^{(k,\ell,\pi)} &= \mathbb{E}\left[g_k^\ell(X_k,U_k) \mid X_k = x\right] \\ M_{xy}^{(k,\pi)} &= P(X_{k+1} = y \mid X_k = x) \ . \end{aligned}$$

The final rewards $\varphi^{(\pi)}$ and $\psi^{(\ell,\pi)}$ are defined similarly. In the sequel, we shall assume that the sets $\mathcal{C}(x)$ are finite, so that the above integrals are replaced by sums.

Definition 7.6. Let $\sigma = (\sigma_k)_{k \ge 0} \in \Sigma^{\mathrm{R}}$ be any random strategy, and $(X,U) := (X_k, U_k)_{k \ge 0}$ be the process induced by σ. The probability law of (X_k, U_k) on $\mathcal{E} \times \mathcal{C}$ is the *occupation measure* of the process at time k. When the sets $\mathcal{C}(x)$ are finite, it is simply written as a map $f^{(k,\sigma)}$ on $\mathcal{E} \times \mathcal{C}$, such that

$$f^{(k,\sigma)}(x,u) = P(X_k = x, U_k = u), \quad x \in \mathcal{E},\ u \in U \ .$$

The occupation measure f^k at time k associated to a Markovian strategy π satisfies, for all $x \in \mathcal{E},\ u \in \mathcal{C}$,

$$f^{(k,\pi)}(x,u) = \pi_k(u \mid x) P(X_k = x) \ . \tag{7.3}$$

Then, in that case, we can recover the law of the process X_k and the strategy π_k at time k from f^k by:

$$P(X_k = x) = p^{(k)}(x) := \sum_{u \in \mathcal{C}} f^{(k,\pi)}(x,u) \tag{7.4a}$$

$$\pi_k(u \mid x) = \frac{f^{(k,\pi)}(x,u)}{p^{(k)}(x)} \ . \tag{7.4b}$$

Proposition 7.7. *For all policies $\pi \in \Pi^{\mathrm{R}}$, the occupation measure satisfies the constraints:*

$$\sum_{u' \in \mathcal{C}} f^{(k+1,\pi)}(y,u') = \sum_{x \in \mathcal{E}} \sum_{u \in \mathcal{C}} M_{xy}^{(k,u)} f^{(k,\pi)}(x,u), \quad \forall y \in \mathcal{E},\ k \geq 0 \ .$$

Moreover, it satisfies the recurrence:

$$f^{(k+1,\pi)}(y,u') = \sum_{x \in \mathcal{E}} \sum_{u \in \mathcal{C}} \Big(\pi_{k+1}(u' \mid y) M_{xy}^{(k,u)} f^{(k,\pi)}(x,u) \Big), \quad \forall y \in \mathcal{E},\ u' \in \mathcal{C} \ , \tag{7.5a}$$

with the constraint on the initial condition

$$f^{(0,\pi)}(x,u') = \pi_0(u' \mid x) \sum_{u \in \mathcal{C}} f^{(0,\pi)}(x,u) \ . \tag{7.5b}$$

Proof. Given $\pi \in \Pi^{\mathrm{R}}$, X_k is a Markov chain with transition probability matrix $M^{(k,\pi)}$ at time k. So the law of X_k can be computed using Fokker-Plank equation:

$$\begin{aligned} P(X_{k+1} = y) &= \sum_{x \in \mathcal{E}} P(X_k = x) M_{xy}^{(k,\pi)} \\ &= \sum_{x \in \mathcal{E}} P(X_k = x) \left(\sum_{u \in \mathcal{C}} M_{xy}^{(k,u)} \pi_k(u \mid x) \right) \end{aligned}$$

Then, using (7.4a) and (7.3), we obtain

$$\sum_{u \in \mathcal{C}} f^{(k+1,\pi)}(y,u) = P(X_{k+1} = y) = \sum_{x \in \mathcal{E}} \sum_{u \in \mathcal{C}} M_{xy}^{(k,u)} f^{(k,\pi)}(x,u) \ .$$

We also obtain:

$$f^{(k+1,\pi)}(y,u') = \sum_{x \in \mathcal{E}} \sum_{u \in \mathcal{C}} \Big(\pi_{k+1}(u' \mid y) M_{xy}^{(k,u)} f^{(k,\pi)}(x,u) \Big) \ ,$$

which can also be obtained as the Fokker-Plank equation for the Markov chain $(X_k, U_k)_{k \geq 0}$, since

$$\pi_{k+1}(u' \mid y) M_{xy}^{(k,u)} = P(U_{k+1} = u', X_{k+1} = y \mid U_k = u, X_k = x) \ .$$

□

Proposition 7.8. *Let $f^{(k)} : \mathcal{E} \times \mathcal{C} \to \mathbb{R}_+$ be functions satisfying the constraints in Proposition 7.7, that is:*

$$\sum_{u' \in \mathcal{C}} f^{(k+1)}(y, u') = \sum_{x \in \mathcal{E}} \sum_{u \in \mathcal{C}} M_{xy}^{(k,u)} f^{(k)}(x, u), \quad \forall y \in \mathcal{E} \ , \tag{7.6}$$

together with the constraints

$$\sum_{u' \in \mathcal{C}} f^{(0)}(x, u') = p^{(0)}(x), \quad \text{and } f^{(k)}(x, u) = 0 \quad \forall u \notin \mathcal{C}_k(x) \ . \tag{7.7}$$

Let π_k be defined as in (7.4), that is:

$$\pi_k(u \mid x) = \frac{f^{(k)}(x, u)}{q^{(k)}(x)}, \quad \text{with } q^{(k)}(x) := \sum_{u \in \mathcal{C}} f^{(k)}(x, u) \ .$$

Then, $\pi = (\pi_k)_{k \geq 0} \in \Pi^{\mathrm{R}}$ and the process $(X_k, U_k)_{k \geq 0}$ associated to the policy π, and the initial law $p^{(0)}$, satisfies

$$f^{(k)}(x, u) = P(X_k = x, U_k = u) \ .$$

Proof. By definition, $\pi_k(\cdot \mid x)$ is a probability on $\mathcal{C}$. Moreover, by the constraints on $f^{(k)}$, the support of $\pi_k(\cdot \mid x)$ is included in $\mathcal{C}_k(x)$.

Let $(X_k, U_k)_{k \geq 0}$ be the process associated to the policy π, and $f^{(k,\pi)}$ be defined as above that is $f^{(k,\pi)}(x, u) = P(X_k = x, U_k = u)$. We need to prove that $f^{(k)} = f^{(k,\pi)}$.

Using the constraints on the functions $f^{(k)}$, we deduce that they satisfy the recurrence:

$$f^{(k+1)}(y, u') = \sum_{x \in \mathcal{E}} \sum_{u \in \mathcal{C}} \Big(\pi_{k+1}(u' \mid y) M_{xy}^{(k,u)} f^{(k)}(x, u) \Big), \quad \forall y \in \mathcal{E}, \ u' \in \mathcal{C} \ ,$$

which is the same as the one of $f^{(k,\pi)}$, see (7.5).

Moreover $f^{(0)}(x, u) = q^{(0)}(x) \pi_0(u \mid x)$, and by the constraints on $f^{(0)}$, we get that $q^{(0)} = p^{(0)}$. Hence $f^{(0)} = f^{(0,\pi)}$, and by induction $f^{(k)} = f^{(k,\pi)}$, for all $k \geq 0$. □

Remark 7.9. In Proposition 7.8, we did not add the constraint that $f^{(k)}$ are probabilities, since this is deduced from the constraints and the properties of $p^{(0)}$ and M.

Corollary 7.10. *Let $f^{(k)} : \mathcal{E} \times \mathcal{C} \to \mathbb{R}_+$, and π_k be as in Proposition 7.8. We have*

$$J^{(T,\pi)} = \widetilde{J}^{(T)}(f), \quad \text{and } G^{(\ell,T,\pi)} = \widetilde{G}^{(\ell,T)}(f)$$

with

$$\widetilde{J}^{(T)}(f) := \left(\sum_{k=0}^{T-1} \sum_{x \in \mathcal{E}, u \in \mathcal{C}} r_k(x, u) f^{(k)}(x, u) \right) + \sum_{x \in \mathcal{E}, u \in \mathcal{C}} \varphi(x) f^{(T)}(x, u) \tag{7.8a}$$

$$\widetilde{G}^{(\ell,T)}(f) := \left(\sum_{k=0}^{T-1} \sum_{x \in \mathcal{E}, u \in \mathcal{C}} g_k^{\ell}(x, u) f^{(k)}(x, u) \right) + \sum_{x \in \mathcal{E}, u \in \mathcal{C}} \psi^{\ell}(x) f^{(T)}(x, u) \ . \tag{7.8b}$$

Theorem 7.11. *The value of the* constrained Markov decision problem *with finite horizon over the set of* Markovian strategies *is equal to the value of the following* Linear Program*:*

$$\max_f \widetilde{J}^{(T)}(f) \text{ under the constraints } \widetilde{G}^{(\ell,T)}(f) \leq h_\ell \quad \forall 1 \leq \ell \leq L \ ,$$

where the maximization is done over the set of sequences $f = (f^{(k)})_{k\geq 0}$ *of functions* $f^{(k)} : \mathcal{E} \times \mathcal{C} \to \mathbb{R}_+$ *satisfying the linear constraints* (7.6) *and* (7.7)*, and* $\widetilde{J}^{(T)}$ *and* $\widetilde{G}^{(\ell,T)}$ *are as in* (7.8)*.*

Proof. Let $\tilde{v}$ be the value of the Linear Program and v be the value of the constrained Markov decision problem with finite horizon over the set of random feedback strategies.

For any $\pi \in \Pi^{\mathrm{R}}$, the sequence $f = (f^{(k,\pi)})_{k\geq 0}$ of occupation measures satisfies all the properties of an argument f of the Linear Program, that is the linear constraints (7.6) and (7.7), and since $\widetilde{J}^{(T)}(f) = J^{(T,\pi)}$ and $\widetilde{G}^{(\ell,T)}(f) = G^{(\ell,T,\pi)}$, we get that $v \leq \tilde{v}$.

Conversely, for any f satisfying the constraints (7.6) and (7.7), Proposition 7.8 constructs $\pi \in \Pi^{\mathrm{R}}$ such that $f^{(k)} = f^{(k,\pi)}$, for all $k \geq 0$. Then, $\tilde{v} \leq v$. □

Theorem 7.12. *The value of the constrained Markov decision problem with finite horizon over the set of* all random strategies *is equal to the value of the same problem over* random Markov strategies *and thus to the value of the* Linear Program*:*

$$\max_f \widetilde{J}^{(T)}(f) \text{ under the constraints } \widetilde{G}^{(\ell,T)}(f) \leq h_\ell \quad \forall 1 \leq \ell \leq L \ ,$$

where the maximization is done over the set of sequences $f = (f^{(k)})_{k\geq 0}$ *of functions* $f^{(k)} : \mathcal{E} \times \mathcal{C} \to \mathbb{R}_+$ *satisfying the linear constraints* (7.6) *and* (7.7)*, and* $\widetilde{J}^{(T)}$ *and* $\widetilde{G}^{(\ell,T)}$ *are as in* (7.8)*.*

This result follows from the following one.

Proposition 7.13. *For all strategies* $\sigma \in \Sigma^{\mathrm{R}}$*, the occupation measure* $f^{(k)} = f^{(k,\sigma)}$ *satisfies the linear constraints* (7.6) *and* (7.7)*, and*

$$J^{(T,\sigma)} = \widetilde{J}^{(T)}(f), \quad \text{and } G^{(\ell,T,\sigma)} = \widetilde{G}^{(\ell,T)}(f) \ .$$

Hence, there exists $\pi \in \Pi^{\mathrm{R}}$ *such that*

$$J^{(T,\sigma)} = J^{(T,\pi)}, \quad \text{and } G^{(\ell,T,\sigma)} = G^{(\ell,T,\pi)} \ .$$

Proof. The proof in Proposition 7.7 was based on Fokker-Plank equation for X_k, but the constraint may be derived from the Fokker-Plank equation for (X_k, U_k), as follows. Moreover, Fokker-Plank equation holds even if (X_k, U_k) is not a Markov chain, so for the process (X_k, U_k) induced by a general random strategy $\sigma \in \Sigma^{\mathrm{R}}$.

Denote $f^{(k)} = f^{(k,\sigma)}$. The Fokker-Plank equation writes

$$f^{(k+1)}(y, u') = \sum_{x\in\mathcal{E}} \sum_{u\in\mathcal{C}} P(U_{k+1} = u', X_{k+1} = y \mid U_k = u, X_k = x) f^{(k)}(x, u) \ .$$

Taking the sum over u', we obtain

$$\sum_{u'\in\mathcal{C}} f^{(k+1)}(y, u') = \sum_{x\in\mathcal{E}} \sum_{u\in\mathcal{C}} P(X_{k+1} = y \mid U_k = u, X_k = x) f^{(k)}(x, u) \ ,$$

and since $P(X_{k+1} = y \mid U_k = u, X_k = x) = M_{xy}^{(k,u)}$, we get (7.6).

The first constraint in (7.7) follows from the fact that $f^{(0)}(x, y) = P(X_0 = x, U_0 = u)$ and the second constraint in (7.7) is by definition of the supports of strategies. □

7.3 Constrained MDP with infinite horizon

Assume given a stationary Markov decision process as in Definition 3.1, that is the following parameters:

- a finite or discrete *state space* $\mathcal{E}$;
- an *action space* $\mathcal{C}$;
- for all $x \in \mathcal{E}$, the subset $\mathcal{C}(x) \subset \mathcal{C}$ of all possible actions at any time k, when the state is equal to x;
- the set $\mathcal{A} := \{(x,u) \mid x \in \mathcal{E},\ u \in \mathcal{C}(x)\}$ of all possibles couples (state, action) (at any time k);
- an initial probability $p^{(0)} \in \Delta_{\mathcal{E}}$ on $\mathcal{E}$, or an initial state $x_0 \in \mathcal{E}$;
- for all $x \in \mathcal{E}$ and $u \in \mathcal{C}(x)$, a probability row vector $M_x^{(u)}$ over $\mathcal{E}$, the entries of which will be denoted $(M_{xy}^{(u)})_{y\in\mathcal{E}}$.

Consider the following (stationary) parameters:

- the *instantaneous rewards* (at any time k), are maps r and g^ℓ, $1 \leq \ell \leq L$, $\mathcal{A} \to \mathbb{R}$;
- a (fixed) *discount factor* $\alpha \in [0,1)$.

For all strategies $\sigma = (\sigma_k)_{k\geq 0}$ in Σ^{R}, consider the *discounted total additive payoffs* with infinite horizon:

$$J_\alpha^{(\sigma)} := J_\alpha(X;U) := \mathbb{E}\left[\sum_{k=0}^{\infty} \alpha^k r(X_k,U_k)\right] \tag{7.9a}$$

$$G_\alpha^{(\ell,\sigma)} := J_\alpha^\ell(X;U) := \mathbb{E}\left[\sum_{k=0}^{\infty} \alpha^k g^\ell(X_k,U_k)\right] \ , \tag{7.9b}$$

where $(X,U) := (X_k,U_k)_{k\geq 0}$ is the process induced by σ (as in Definition 3.2 or Definition 3.3).

Proposition 7.14. *For all strategies $\sigma \in \Sigma^{\mathrm{R}}$, and all $\alpha \in [0,1)$, denote, for $x \in \mathcal{E}$ and $u \in \mathcal{C}(x)$,*

$$f^{(\sigma)}(x,u) = \sum_{k=0}^{\infty} \alpha^k f^{(k,\sigma)} \ .$$

This function satisfies the constraints:

$$\sum_{u'\in\mathcal{C}} f^{(\sigma)}(y,u') = p^{(0)}(y) + \alpha \sum_{x\in\mathcal{E}} \sum_{u\in\mathcal{C}} M_{xy}^{(u)} f^{(\sigma)}(x,u), \quad \forall y \in \mathcal{E} \ .$$

Moreover, if $\sigma = \pi$ is a stationary Markovian strategy, we have:

$$f^{(\pi)}(y,u') = \pi(u' \mid y)\left(\sum_{u\in\mathcal{C}} f^{(\pi)}(y,u)\right), \quad \forall y \in \mathcal{E},\ u' \in \mathcal{C} \ . \tag{7.10}$$

Proof. Multiplying the constraint in Proposition 7.7 (which is also true for any strategy $\sigma \in \Sigma^{\mathrm{R}}$) by α^{k+1}, and summing all the equations for $k \geq 0$, we obtain

$$\sum_{u' \in \mathcal{C}} (f^{(\sigma)}(y,u') - f^{(0,\sigma)}(y,u')) = \alpha \sum_{x \in \mathcal{E}} \sum_{u \in \mathcal{C}} M_{xy}^{(k,u)} f^{(\pi)}(x,u), \quad \forall y \in \mathcal{E},\ k \geq 0 \ .$$

Using (7.4a), we deduce the constraint of the proposition. The last equation is also obtained by summing all the equations in (7.3) after multiplication by α^k. □

Proposition 7.15. *Let $f : \mathcal{E} \times \mathcal{C} \to \mathbb{R}_+$ satisfies the constraint of Proposition 7.14, that is*

$$\sum_{u' \in \mathcal{C}} f(y,u') = p^{(0)}(y) + \alpha \sum_{x \in \mathcal{E}} \sum_{u \in \mathcal{C}} M_{xy}^{(u)} f(x,u), \quad \forall y \in \mathcal{E} \ , \tag{7.11}$$

together with the constraints

$$f(x,u) = 0 \quad \forall u \notin \mathcal{C}(x) \ . \tag{7.12}$$

Let π be defined by:

$$\pi(u \mid x) = \frac{f(x,u)}{q(x)}, \quad \text{with } q(x) := \sum_{u \in \mathcal{C}} f(x,u) \ .$$

Then, π is a stationary Markovian strategy and the process $(X_k, U_k)_{k \geq 0}$ associated to the policy π and the initial law $p^{(0)}$ satisfies

$$f(x,u) = \sum_{k=0}^{\infty} \alpha^k P(X_k = x, U_k = u) \ .$$

Corollary 7.16. *Let $f : \mathcal{E} \times \mathcal{C} \to \mathbb{R}_+$ and π be as in Proposition 7.15. We have*

$$J_\alpha^{(\pi)} = \widetilde{J}(f), \quad \text{and} \quad G_\alpha^{(\ell,\pi)} = \widetilde{G}^{(\ell)}(f)$$

with

$$\widetilde{J}(f) := \sum_{x \in \mathcal{E}, u \in \mathcal{C}} (r(x,u) f(x,u)) \tag{7.13a}$$

$$\widetilde{G}^{(\ell)}(f) := \sum_{x \in \mathcal{E}, u \in \mathcal{C}} \Big(g^\ell(x,u) f(x,u) \Big) \ . \tag{7.13b}$$

Theorem 7.17. *The value of the constrained Markov decision problem with discounted infinite horizon criteria over the set of* all random strategies *is equal to the value of the same problem over* random stationary Markov strategies *and to the value of the* Linear Program*:*

$$\max_f \widetilde{J}(f) \text{ under the constraints } \widetilde{G}^{(\ell)}(f) \leq h_\ell \quad \forall 1 \leq \ell \leq L \ ,$$

where the maximization is done over the set of functions $f : \mathcal{E} \times \mathcal{C} \to \mathbb{R}_+$ satisfying the linear constraints (7.11) *and* (7.12)*, and $\widetilde{J}$ and $\widetilde{G}^{(\ell)}$ are as in* (7.13)*.*

7.4 Constrained MDP with long run time average payoff

Similarly, passing to the limit when $\alpha \to 1^-$, we can obtain, assuming some ergodicity:

Theorem 7.18. *The value of the constrained Markov decision problem with long run time average payoff over the set of* all random strategies *is equal to the value of the same problem over* random stationary Markov strategies *and to the value of the* Linear Program*:*

$$\max_f \widetilde{J}(f) \text{ under the constraints } \widetilde{G}^{(\ell)}(f) \le h_\ell \quad \forall 1 \le \ell \le L \ ,$$

where the maximization is done over the set of functions $f : \mathcal{E} \times \mathcal{C} \to \mathbb{R}_+$ *satisfying the linear constraints* (7.12) *and*

$$\sum_{u' \in \mathcal{C}} f(y, u') = \sum_{x \in \mathcal{E}} \sum_{u \in \mathcal{C}} M_{xy}^{(u)} f(x, u), \quad \forall y \in \mathcal{E} \ , \tag{7.14}$$

and $\widetilde{J}$ *and* $\widetilde{G}^{(\ell)}$ *are as in* (7.13)*.*

(7.14) means that f is *an invariant occupation measure.*

7.5 Complexity

For solving an unconstrainted discounted infinite horizon MDP with finite state and action spaces, with $m = \max_k \mathrm{card}(\mathcal{A}_k)$ and $n = \mathrm{card}(\mathcal{E})$, the complexity of value iterations was:

$$\mathcal{O}\Big(\frac{\log((1-\alpha)\varepsilon/R_{\max})}{\log(\alpha)}\Big) nm \ ,$$

where $R_{\max}$ is a bound on the rewards. It is only pseudo-polynomial because $\log(\alpha)$ is exponential in the number of bit of α. If α is fixed, it is polynomial but not strongly polynomial, since it depends on the number of bit of the rewards.

For a constrained infinite horizon MDP, the size of the variable f of the corresponding Linear Program is in $\mathcal{O}(m)$ with L linear inquality constraints and n linear equality constraints.

One can use the simplex or interior point algorithms.

Ye [CS3] proved that simplex for unconstrained MDP is strongly polynomial when α is fixed.

Interior point is polynomial but not strongly polynomial.

Additional references for this chapter

[CS1] Eitan Altman Constrained Markov Decision Processes. *CRC Press, 1999.* See also https://www-sop.inria.fr/members/Eitan.Altman/TEMP/h.pdf

[CS2] Richard C. Chen and Gilmer L. Blankenship. Dynamic programming equations for discounted constrained stochastic control. *IEEE Trans. Autom. Control*, 49(5):699–709, 2004.

[CS3] Y. Ye. The simplex and policy-iteration methods are strongly polynomial for the Markov decision problem with a fixed discount rate. *Math. Oper. Res.*, 36(4):593–603, 2011.